JRW

ASA-1162

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of

A. SATOYAMA et al

Serial No. 10/766,823

Filed: January 30, 2004

For: STORAGE SYSTEM AND REPLICATION CREATION METHOD THEREOF

TRANSMITTAL OF CERTIFIED PRIORITY DOCUMENT

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Submitted herewith is a certified priority document (JP 2003-403868, filed December 3, 2003) of a corresponding Japanese patent application for the purpose of claiming foreign priority under 35 U.S.C. §119. An indication that this document has been safely received would be appreciated.

Respectfully submitted,

Daniel J. Stanger
Registration No. 32,846
Attorney for Applicants

MATTINGLY, STANGER & MALUR
1800 Diagonal Rd., Suite 370
Alexandria, Virginia 22314
(703) 684-1120
Date: June 22, 2004

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: ２００３年１２月 ３日

出 願 番 号
Application Number: 特願２００３－４０３８６８

[ST. 10/C]: [ＪＰ２００３－４０３８６８]

出 願 人
Applicant(s): 株式会社日立製作所

A. Satoyama et al
10/766,823
filed 1-30-04
703-684-1120
ASA-1162

２００４年 １月２７日

特許庁長官
Commissioner,
Japan Patent Office 今 井 康 夫

出証番号 出証特２００４－３００３０１５

【書類名】　特許願
【整理番号】　K03010501
【あて先】　特許庁長官殿
【国際特許分類】　G06F 12/00
【発明者】
　【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
　【氏名】　里山　愛
【発明者】
　【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
　【氏名】　山本　康友
【発明者】
　【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
　【氏名】　森下　昇
【発明者】
　【住所又は居所】　神奈川県川崎市麻生区王禅寺１０９９番地　株式会社日立製作所システム開発研究所内
　【氏名】　江口　賢哲
【特許出願人】
　【識別番号】　000005108
　【氏名又は名称】　株式会社日立製作所
【代理人】
　【識別番号】　100083552
　【弁理士】
　【氏名又は名称】　秋田　収喜
　【電話番号】　03-3893-6221
【手数料の表示】
　【予納台帳番号】　014579
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　特許請求の範囲　1
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1

【書類名】特許請求の範囲
【請求項１】
複数のディスク装置が接続された制御部を複数個有する記憶装置システムであって，
前記制御部に接続されたディスク装置内のボリュームのデータのレプリケーションを作成するレプリケーション作成部と，
レプリケーション元のボリュームとレプリケーション先のボリュームに関する情報であるペア情報と，
を各制御部毎に備え，
前記複数の制御部のうちの一制御部のレプリケーション作成部は，
同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は，前記ペア情報に，レプリケーション元のボリューム情報と，レプリケーション先のボリューム情報と，を登録し，前記ペア情報に基づいて前記同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成し，
他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は，前記ペア情報に，レプリケーション元のボリューム情報と，前記一制御部におけるレプリケーション先のボリューム情報と，前記他の制御部に関する情報と，を登録し，前記ペア情報に基づいて，レプリケーションを作成するための要求を前記他の制御部へ送信することによりレプリケーションを作成する
ことを特徴とする記憶装置システム。
【請求項２】
請求項１に記載の記憶装置システムであって，
前記各制御部はデータを一時的に記憶するキャッシュを備え，
前記一制御部のレプリケーション作成部は，
他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合に，レプリケーション元のボリュームのデータを，前記一制御部における前記キャッシュ内にレプリケーション先のボリュームのデータ用にコピーし，前記キャッシュ内にコピーされた前記データを，前記他の制御部に送信する
ことを特徴とする記憶装置システム。
【請求項３】
請求項２に記載の記憶装置システムであって，
前記他の制御部は，前記一制御部から受信した前記データを，該制御部のキャッシュに格納した後に，該制御部に接続されたディスク装置内のボリュームに格納することを特徴とする記憶装置システム。
【請求項４】
請求項１に記載の記憶装置システムであって，
前記一制御部におけるレプリケーション先のボリューム情報は，前記一制御部に接続されたディスク装置にボリュームが作成されることがない仮想的なボリューム情報であり，
前記他の制御部に関する情報は，前記他の制御部を識別するための情報と前記他の制御部に接続されたディスク装置内のレプリケーション先のボリュームのボリューム情報である
ことを特徴とする記憶装置システム。
【請求項５】
請求項１に記載の記憶装置システムであって，
前記ペア情報は，レプリケーション元のボリュームとレプリケーション先のボリュームのペアに対して割り当てられた識別子を含み，
一個のレプリケーション元のボリュームに対して一個または複数個の識別子を組むことを特徴とする記憶装置システム。
【請求項６】
請求項１に記載の記憶システムであって，
前記レプリケーション作成部は，

管理端末から入力された情報若しくはホストからのホストコマンドに基づいて，または，自動的に，
前記ペア情報への情報の登録を行うことを特徴とする記憶装置システム。
【請求項７】
請求項１記載の記憶装置システムであって，
通常のリード／ライト要求をスケジュールするための通常リード／ライト処理キューと，
通常のリード／ライト要求よりも処理の優先度が低い要求をスケジュールするための処理優先度低キューと，
を各制御部毎に備え，
前記レプリケーションを作成するための要求を通常リード／ライト要求と同等に処理する場合は，該要求を前記通常リード／ライト処理キューに入れ，
前記レプリケーションを作成するための要求を通常リード／ライト要求より低い優先度で処理する場合は，該要求を前記処理優先度低キューに入れる
ことを特徴とする記憶装置システム。
【請求項８】
請求項１に記載の記憶装置システムであって，
前記レプリケーションを作成するための要求を通常のリード／ライト要求と同等に処理すべきかどうかの処理優先度を設定し，他の制御部へ処理優先度情報を通知する処理優先度設定部を各制御部毎に備え，
前記通知を受けた他の制御部が，前記処理優先度情報により，要求を処理する順番をスケジュールすることを特徴とする記憶装置システム。
【請求項９】
請求項１に記載の記憶装置システムであって，
一制御部が，他の制御部へレプリケーションを作成するための要求を送信する場合に，送信する制御命令にレプリケーション作成処理であることを示す情報を付加して送信し，
前記他の制御部が，前記情報に基づいて前記要求を優先処理するかどうかを決定し，前記要求を処理する順番をスケジュールすることを特徴とする記憶装置システム。
【請求項１０】
請求項１に記載の記憶装置システムであって，
一制御部が，他の制御部へレプリケーションを作成するための要求を送信する場合に，送信する要求命令に優先度を示す情報を付加して送信し，
前記他の制御部が，前記情報に基づいて前記要求を優先処理するかどうかを決定し，前記要求を処理する順番をスケジュールすることを特徴とする記憶装置システム。
【請求項１１】
複数のディスク装置が接続された制御部を複数個有する記憶装置システムにおけるレプリケーション作成方法であって，
前記記憶装置システムは，
前記制御部に接続されたディスク装置内のボリュームのデータのレプリケーションを作成するレプリケーション作成部と，
レプリケーション元のボリュームとレプリケーション先のボリュームに関する情報であるペア情報と，
を各制御部毎に備え，
同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は，前記複数の制御部のうちの一制御部のレプリケーション作成部が，前記ペア情報に，レプリケーション元のボリューム情報と，レプリケーション先のボリューム情報と，を登録するステップと，
他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は，前記一制御部のレプリケーション作成部が，前記ペア情報に，レプリケーション元のボリューム情報と，前記一制御部におけるレプリケーション先のボリューム情報と，前

記他の制御部に関する情報と，を登録するステップと，
　同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は，前記一制御部のレプリケーション作成部が，前記ペア情報に基づいて前記同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成するステップと，
　他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は，前記一制御部のレプリケーション作成部が，前記ペア情報に基づいて，レプリケーションを作成するための要求を前記他の制御部へ送信することによりレプリケーションを作成するステップと
を有することを特徴とするレプリケーション作成方法。
【請求項１２】
　請求項１１に記載のレプリケーション作成方法であって，
　前記記憶装置システムの前記各制御部はデータを一時的に記憶するキャッシュを備え，
　他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合に，前記一制御部のレプリケーション作成部が，
レプリケーション元のボリュームのデータを，前記一制御部における前記キャッシュ内にレプリケーション先のボリュームのデータ用にコピーし，前記キャッシュ内にコピーされた前記データを，前記他の制御部に送信する
ことを特徴とするレプリケーション作成方法。
【請求項１３】
　請求項１２に記載のレプリケーション作成方法であって，
　前記他の制御部は，前記一制御部から受信した前記データを，該制御部のキャッシュに格納した後に，該制御部に接続されたディスク装置内のボリュームに格納することを特徴とするレプリケーション作成方法。
【請求項１４】
　請求項１１に記載のレプリケーション作成方法であって，
　前記一制御部におけるレプリケーション先のボリューム情報は，前記一制御部に接続されたディスク装置にボリュームが作成されることがない仮想的なボリューム情報であり，
　前記他の制御部に関する情報は，前記他の制御部を識別するための情報と前記他の制御部に接続されたディスク装置内のレプリケーション先のボリュームのボリューム情報である
ことを特徴とするレプリケーション作成方法。
【請求項１５】
　請求項１１に記載のレプリケーション作成方法であって，
　前記ペア情報は，レプリケーション元のボリュームとレプリケーション先のボリュームのペアに対して割り当てられた識別子を含み，
　一個のレプリケーション元のボリュームに対して一個または複数個の識別子を組むことを特徴とするレプリケーション作成方法。
【請求項１６】
　請求項１１に記載のレプリケーション作成方法であって，
　前記レプリケーション作成部は，
　管理端末から入力された情報若しくはホストからのホストコマンドに基づいて，または，自動的に，
　前記ペア情報への情報の登録を行うことを特徴とするレプリケーション作成方法。
【請求項１７】
　請求項１１記載のレプリケーション作成方法であって，
　前記記憶装置システムは，
　通常のリード／ライト要求をスケジュールするための通常リード／ライト処理キューと，
　通常のリード／ライト要求よりも処理の優先度が低い要求をスケジュールするための処理優先度低キューと，

を各制御部毎に備え，
　前記レプリケーションを作成するための要求を通常リード／ライト要求と同等に処理する場合は，該要求を前記通常リード／ライト処理キューに入れ，
　前記レプリケーションを作成するための要求を通常リード／ライト要求より低い優先度で処理する場合は，該要求を前記処理優先度低キューに入れる
ことを特徴とするレプリケーション方法。
【請求項１８】
　請求項１１に記載のレプリケーション作成方法であって，
　前記記憶装置システムは処理優先度設定部を各制御部毎に備え，
　一制御部の処理優先度設定部が，前記レプリケーションを作成するための要求を通常のリード／ライト要求と同等に処理すべきかどうかの処理優先度を設定し，他の制御部へ処理優先度情報を通知し，
　前記通知を受けた他の制御部が，前記処理優先度情報により，要求を処理する順番をスケジュールすることを特徴とするレプリケーション作成方法。
【請求項１９】
　請求項１１に記載のレプリケーション作成方法であって，
　一制御部が，他の制御部へレプリケーションを作成するための要求を送信する場合に，送信する制御命令にレプリケーション作成処理であることを示す情報を付加して送信し，
　前記他の制御部が，前記情報に基づいて前記要求を優先処理するかどうかを決定し，前記要求を処理する順番をスケジュールすることを特徴とするレプリケーション作成方法。
【請求項２０】
　請求項１１に記載のレプリケーション作成方法であって，
　一制御部が，他の制御部へレプリケーションを作成するための要求を送信する場合に，送信する要求命令に優先度を示す情報を付加して送信し，
　前記他の制御部が，前記情報に基づいて前記要求を優先処理するかどうかを決定し，前記要求を処理する順番をスケジュールすることを特徴とするレプリケーション作成方法。

【書類名】明細書
【発明の名称】記憶装置システムおよびそのレプリケーション作成方法
【技術分野】
【０００１】
本発明は，複数のディスク装置が接続された制御部を複数個有する記憶装置システムおよびそのレプリケーション作成方法に関する。
【背景技術】
【０００２】
近年，企業が保有する記憶装置に格納されたデータの複製を，別の記憶装置に作成する処理（以下，「バックアップ」と称する）に係る時間を短縮したいという要求が高まっている。これは，企業が保有する情報量の増加に伴ってバックアップに係る時間が益々増加する一方で，企業の業務時間の延長により，バックアップの処理に割り当てられる時間が短くなっていることが背景にある。
【０００３】
特許文献１，特許文献２に開示されたように，企業における日常の業務を停止せずに，記憶装置に格納されたデータをバックアップする技術として，スナップショットが提案されている。スナップショットとは，記憶装置と接続される計算機を介さずに，記憶装置が有する記憶領域の特定の時点におけるコピーを記憶装置に作成する機能である。この機能を利用して，ユーザは，元の記憶領域を業務で使用し，コピーされた記憶領域に格納されたデータをバックアップに使用する。
【０００４】
【特許文献１】特開平７－２１０４３９号公報
【特許文献２】特開２００１－３１８８３３号公報
【発明の開示】
【発明が解決しようとする課題】
【０００５】
ネットワークに接続される記憶装置のスケーラビリティを高めるための技術として，クラスタ構成記憶装置システムが考えられる。クラスタ構成記憶装置システムとは，ディスクアレイ装置のような従来の記憶装置システムを１クラスタとし，一つの記憶装置システムを複数のクラスタから構成した記憶装置システムである。
【０００６】
従来，クラスタ構成記憶装置システムにおいてスナップショットを行うことを示唆した文献はない。また，クラスタ構成記憶装置システムと従来のスナップショット技術を単純に組み合わせた場合，１クラスタ内に限って，記憶領域のコピーを実行する技術となる。
【０００７】
しかし，異なるクラスタ間で記憶領域のコピーを作成することが出来ないと，１つのクラスタ構成記憶装置システム内で，記憶領域のコピー先として使用できる記憶領域と使用できない記憶領域ができてしまい，クラスタ構成記憶装置システムの本来の目的であるスケーラビリティが損なわれる。
【０００８】
また，クラスタ構成記憶装置システムにおいて，クラスタ間にまたがった論理ボリューム（以下，「ボリューム」という。）のコピーを作成する場合，つまり，コピー元ボリュームとコピー先ボリュームが異なるクラスタにある場合，コピー元のボリュームがあるクラスタ（以下正クラスタと呼ぶ）は，コピー先のボリュームがあるクラスタ（以下副クラスタと呼ぶ）内の共有メモリを参照できないため，副クラスタの負荷状態について認識できない。このため，コピー元と同一クラスタ内でコピー先のボリュームを選択するという，ユーザの制約が生じるため，従来と装置構成が異なることによるユーザの使い勝手が変更になる。
【０００９】
本発明の目的は，複数のディスク装置が接続された制御部を複数個有する記憶装置シス

テムにおいて，同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合だけでなく，異なった制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合も，異なった制御部であることを意識することなく記憶領域のコピーを作成することができる記憶装置システムを提供できるようにすることである。

【課題を解決するための手段】

【００１０】

前記目的を達成するために，本発明の記憶装置システムは，複数のディスク装置が接続された制御部を複数個有する記憶装置システムであって，前記制御部に接続されたディスク装置内のボリュームのデータのレプリケーションを作成するレプリケーション作成部と，レプリケーション元のボリュームとレプリケーション先のボリュームに関する情報であるペア情報と，を各制御部毎に備える。そして，前記複数の制御部のうちの一制御部のレプリケーション作成部は，同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は，前記ペア情報に，レプリケーション元のボリューム情報と，レプリケーション先のボリューム情報と，を登録し，前記ペア情報に基づいて前記同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成し，他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は，前記ペア情報に，レプリケーション元のボリューム情報と，前記一制御部におけるレプリケーション先のボリューム情報と，前記他の制御部に関する情報と，を登録し，前記ペア情報に基づいて，レプリケーションを作成するための要求を前記他の制御部へ送信することによりレプリケーションを作成する。

【発明の効果】

【００１１】

本発明により，複数のディスク装置が接続された制御部を複数個有する記憶装置システムにおいて，同一制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合だけでなく，異なった制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合も，異なった制御部であることを意識することなく記憶領域のコピーを作成することができる。

【発明を実施するための最良の形態】

【００１２】

図１に本発明の一実施例のクラスタ構成記憶装置システムを含む計算機システムの一例を示す。

第一の記憶装置システム７０Ａは，ネットワーク６０を介してホスト１０，１１と接続されている。管理端末８０は，管理用ネットワーク６２を介して，第二の記憶装置システム７０Ｂは，ネットワーク６１を介して，第一の記憶装置システム７０Ａに接続される。ここでは，第二の記憶装置システム７０Ｂが接続されている構成を図示するが，７０Ｂが接続されていない構成でもよい。

【００１３】

記憶装置システム７０Ａは，プロトコル変換アダプタ４０を有する。プロトコル変換アダプタ４０は，記憶制御装置２０と独立したチャネル接続部分であり，ＬＡＮ（Local Area Network），公衆回線，専用回線，ＥＳＣＯＮ（Enterprise Systems Connection）に準拠したプロトコルなどを扱う。複数のプロトコル変換アダプタ４０と複数の記憶制御装置２０はネットワーク６３を介して接続されている。

【００１４】

プロトコル変換アダプタ４０は，ホスト１０，１１から入出力要求を受け取ると，コマンドを解析してプロトコル変換し，前記コマンドが要求するデータが格納されているＬＵ（Logical Unit）がどの記憶制御装置２０の配下に管理されているか，記憶装置システム７０Ｂに管理されているかを判断し，判断した箇所に当該要求を送信する。前記どの装置で管理されているかは，ネットワーク６３で接続されたプロセッサ５０内のメモリ上に格納された構成情報テーブル５１０を参照することで判断する。

【００１５】
管理端末８０は，ネットワーク６２を介して，第一の記憶装置システム７０Ａを認識する。また，管理端末８０を，第一の記憶装置システム７０Ａと専用線で直接接続する構成もある。

【００１６】
記憶制御装置２０は，ＣＰＵ２１，メモリ２２，ホスト１０，１１等からの入出力データを一時的に格納するキャッシュ２３，ネットワーク６３と接続する部分であるハブ２４，記憶装置群３１との間のデータの送受信を制御する記憶装置Ｉ／Ｆ２５を有し，これらは，内部バスで相互に接続される。

【００１７】
図２に，メモリ２２の一実施例の構成図を示す。メモリ２２には，図２に示すように，ＣＰＵ２１で実行される種々のプログラムが格納される。具体的には，記憶装置システム７０の動作を制御するためのＲＡＩＤ制御プログラム２００及び記憶装置システム７０の構成を管理するための管理エージェント２１０である。又，メモリ２２には，各種管理情報も格納される。具体的には，データのコピー元（レプリケーション元）とコピー先（レプリケーション先）に関する情報を記録するペア情報テーブル２２０，ボリューム情報テーブル２３０，処理優先度情報ビットマップ２５０，差分ビットマップ２４０，処理キュー２６０，記憶装置システム７０Ｂが記憶装置システム７０Ａに対し自身のＬＵを記憶装置システム７０ＡのＬＵとして提供するための構成情報テーブル（図示せず）を有する。前記２２０～２５０及び構成情報テーブルはプロセッサ５０においてもよい。

【００１８】
ＲＡＩＤ（Redundant Array of Inexpensive Disks）制御プログラム２００は，記憶装置群３１にコマンドを発行する部分（不図示）を有する。ＲＡＩＤ制御プログラム２００内には，サブプログラムとして，記憶装置システム７０内でデータのレプリケーション（コピー）を作成するためのレプリケーション作成プログラム２０１と処理優先度設定プログラム２０２がある。ＲＡＩＤ制御プログラム２００，レプリケーション作成プログラム２０１，処理優先度設定プログラム２０２はＣＰＵ２１（図１参照）で実行され，それにより，記憶制御装置２０のＲＡＩＤ制御部，レプリケーション作成部，処理優先度設定部が構成される。また，ＲＡＩＤ制御プログラム２００，レプリケーション作成プログラム２０１，処理優先度設定プログラム２０２の一部または全部をハードウェアで構成してもよい。尚，データのレプリケーションを実施する際，同期（データのコピーの完了を待って，上位装置に完了を報告する），非同期（データのコピーの完了を待たずに，上位装置に完了を報告する）といったバリエーションがあるが，本実施形態では，特に区別しない。

管理エージェント２１０は，管理端末８０からの入力を受けて，情報を設定したり，情報を管理端末８０へ出力するためのプログラムである。

【００１９】
図３に，管理端末８０の一実施例の構成図を示す。管理端末８０は，ＣＰＵ８１，主記憶８２，入力部（キーボード装置等）８３，出力部（ディスプレイ装置等）８４，ネットワーク６２と接続するための管理Ｉ／Ｆ８５及び，記憶部８６を有し，これらは，内部バスで相互に接続される。記憶部８６には，ＣＰＵ８１で実行されるプログラムである処理優先度設定プログラム２０２が格納される。

ホスト１０は，例えば，パソコンやワークステーション，汎用コンピュータなどであり，外部に接続するためのＦＣインタフェースであるＨＢＡ（Host Bus Adaptor）（不図示）を備える。ＨＢＡにもＷＷＮ（World Wide Name）が付与されている。

【００２０】
図４は，ペア情報テーブル２２０の一例を示す図である。本テーブルは，記憶装置システム７０内の，コピーされたデータを保持するボリュームの対（以下「ペア」）を管理するための情報であり，ペア番号２２１，正ボリューム情報２２２，副ボリューム情報２２３～２２５，ペア状態２２６がある。特に，記憶制御装置２０Ａのペア及び記憶制御装置

２０Ａにコピー元があるペアの情報は，記憶制御装置２０Ａ内のメモリ２２におく。
ペア番号２２１は，ペアに対して任意に割り当てられた識別子を示す。
正ボリューム情報２２２は，識別子が与えられたペアのうち，レプリケーション元（コピー元）のボリュームである正ボリュームに割り振られたボリューム番号を示す。

【００２１】

副ボリューム情報２２３～２２５は，識別子が与えられたペアのうち，レプリケーション先（コピー先）のボリュームである副ボリュームに関する情報である。同一記憶制御装置内のペアの場合は，副ボリューム番号を２２３に登録する。異記憶制御装置間のペアの場合は，同一記憶制御装置内に仮想化したボリューム番号を２２３に登録し，実際にデータの格納される副ボリュームの記憶制御装置に関する情報，例えば記憶制御装置番号を２２４に，ボリューム番号を２２５に登録する。同一記憶制御装置内に仮想化したボリューム番号は，その同一記憶制御装置におけるボリューム番号であるが，その同一記憶制御装置に接続されたディスク装置にボリュームが作成されることがない仮想的なボリューム番号である。

【００２２】

ペア状態２２６は，ペアの現在の状態を示す。例えば，ペアの各々のボリュームに格納されているデータの同期が取られ，各々に格納されたデータの内容が一致している状態(以下「Ｐａｉｒ状態」と称する)又は，ペア間でデータの同期が取られていない状態（以下「Ｓｐｌｉｔ状態」と称する）などがある。

【００２３】

記憶装置システム７０Ａは，例えば，Ｐａｉｒ状態にあったペアを，任意の時間においてＳｐｌｉｔ状態に変更する。このとき，副ボリュームには，任意の時間におけるペアの正ボリュームが有するデータが保存される（このような処理を「スナップショットを取る」という）。その後，ホスト１０が副ボリュームからデータを読み出して他の記憶装置（例えばテープ装置等）に書き込むことで，スナップショットを取った時間におけるペアに格納されたデータのバックアップをとることができる。尚，スナップショットを取った後の副ボリュームそのものをデータのバックアップとして保存しておいてもよい。

【００２４】

図１０は，ボリューム情報テーブル２３０の一例を示す図である。本テーブルは，記憶制御装置２０Ａ配下のボリュームを管理する情報を登録し，記憶制御装置２０Ａ内のメモリに格納する。
ボリューム番号２３１は，ボリュームに対して割り当てられた識別子である。ボリューム番号０は，ボリューム番号０に対してペアを３個組んだ例である。一般には，１個のボリュームに対して１個または複数個のペアを組むことができる。
正／副２３２は，ボリュームがペアの正の役割をしているか，副の役割をしているかを示す。

【００２５】

相手ボリューム情報の２３３～２３５は，ペアを組んでいる場合の相手のボリューム情報である。同一記憶制御装置内のペアの場合は，副ボリューム番号を２３３に登録する。異記憶制御装置間のペアの場合は，同一記憶制御装置内に仮想化したボリューム番号を２３３に登録し，実際にデータの格納される副ボリュームの記憶制御装置番号を２３４に，ボリューム番号を２３５に登録する。ボリューム使用中２３６は，そのボリュームが使用中か空いているかを示す情報である。

【００２６】

図１０は記憶制御装置０番のボリューム情報であるとする。ボリューム０番は３つのペアを作成している。第１のペアは，記憶制御装置０番のボリューム１０２４番を仮想化副ボリュームとし，実際のデータの格納されている副ボリュームは記憶制御装置１番のボリューム２０番であることを示す。第２のペアは同一記憶制御装置内のボリューム１５８番であることを示す。ボリューム１番は，ペアの副ボリュームとして使用されており，正ボリュームは記憶制御装置３番のボリューム３７８３番であることを示す。

【００２７】
図１１は，差分ビットマップ２４０の一例を示す図である。１つのペアに対して同じサイズのビットマップを２つ用意する。差分ビットマップの‘０’はコピーが終了した箇所，‘１’はコピーが終了していない箇所を表している。１ビットに対して予め決めたデータサイズのデータを対応させている。例えば，１ビットに対し６４ＫＢのデータを対応させた場合，６４ＫＢ中の１Ｂでも更新があった場合は，ビットを‘１’にして，内容がコピー先にも反映されるようにする。
【００２８】
図１４は，処理優先度情報ビットマップ２５０の一例を示す図である。本ビットマップは，ビットマップが格納されている記憶制御装置配下のボリュームで，他の記憶制御装置配下のボリュームとレプリケーションのペアを組んでいるものに対する情報を示す。ビットが‘１’の場合，通常のリード／ライト処理より優先度を下げて処理したいことを示す。ビットが‘０’の場合は，通常のリード／ライト処理と同等にスケジュールして処理したいことを示す。図１４のビットマップはボリューム番号順に対応しているとすると，先頭のビットはボリューム番号０の情報となる。このビットマップ例のうち，ビットが‘１’であるのは，４番目，５番目，８番目であることから，ボリューム番号３，４，７の処理は通常リード／ライト処理の要求頻度が低い時に処理して，通常リード／ライト処理の性能に影響を与えないように処理することを示す。本ビットマップ例では‘０’と‘１’で示したが，優先度を数段階にわけてスケジュールさせたい場合，１ボリュームに対するビット数を増やしても良い。
【００２９】
このシステム構成において，記憶制御装置２０Ａにあるボリュームのコピーボリュームを記憶制御装置２０Ｂに作る場合の方式について説明する。
コピーボリュームの作成処理は，レプリケーション作成プログラム２０１が実施する。レプリケーション作成プログラム２０１は，レプリケーションペアの正副ボリュームが同じ記憶制御装置内か，異なる記憶制御装置間かをチェックし，異記憶制御装置間の場合は，本発明の処理を行う。同一記憶制御装置内のペアの場合は，従来の処理を行う。
【００３０】
図５に，図１のクラスタ構成記憶装置システムでレプリケーション作成する際のレプリケーション作成処理の概要のフローチャートを示す。まず，空きボリュームの中からコピー先となる副ボリュームを選択し，正副ボリュームをペアとして，ペア情報テーブル２２０に登録する（ステップ５０１０）。ペアが同一記憶制御装置２０内にあるか否かを判定し（ステップ５０２０），同一記憶制御装置２０内にある場合は，同一記憶制御装置内のレプリケーション作成処理を行い（ステップ５０３０），処理を終了する。もし，異なる記憶制御装置２０間である場合は，異記憶制御装置間レプリケーション作成処理を行い（ステップ５０４０），処理を終了する。
【００３１】
次に，前記ステップ５０３０の同一記憶制御装置内レプリケーション作成処理方式を図６のフローチャートに従って説明する。
正ボリュームの内容を副ボリュームに全コピーする初期コピー処理を開始する。初期コピーは，図２，図１１に示す差分ビットマップ２４０Ｐ１の全ビットを‘１’にする（ステップ６０１０）。処理が順次差分ビットマップ上の‘１’を検出したら（ステップ６０２０），そのビットに対応する箇所のデータがキャッシュにあるか否かの判定を行う（ステップ６０３０）。ステップ６０２０で検出しなければ，ステップ６０７０へ進む。ステップ６０３０でキャッシュになければ正ボリュームからキャッシュにリードする（ステップ６０４０）。キャッシュ内で副ボリュームのデータ用にコピーする（ステップ６０５０）。コピーに伴い，データが正しいか否かを判定するための冗長情報も副ボリューム用に新規に作成しデータへ付随させる。キャッシュに格納したら，差分ビットを‘０’にする（ステップ６０６０）。次のビットがあれば，ステップ６０２０から６０６０を繰り返す（ステップ６０７０）。次のビットが無ければ，本処理は終了する。また，前記の正ボリ

ユームからデータをリードする際に，副ボリューム用の冗長情報を作成し直接副ボリューム用のデータとしてキャッシュに格納する方法もある。一方，非同期でキャッシュ上の副ボリューム用データを副ボリュームに格納する（ステップ６０８０）。

【００３２】

次に，前記ステップ５０４０の異記憶制御装置間レプリケーション作成処理方法を図７，図８を用いて説明する。

図７に，異記憶制御装置間レプリケーションを作成する一実施例を示す。本例では，ボリューム３１１をコピー元としボリューム３１２にコピーしたい。その際，正側ボリューム３１１のある記憶制御装置２０Ａにて，記憶制御装置２０Ｂにある副側ボリューム３１２を記憶制御装置２０Ａ内のボリューム３１３に仮想化する。これにより，ホストは同一記憶制御装置２０Ａ内のペア（ボリューム３１１と３１３がペアでペア番号１）としてレプリケーション作成処理を実行する。一方，記憶制御装置２０Ａは，実際に副ボリュームは記憶制御装置２０Ｂ内にあることを認識し，コピーデータを記憶制御装置２０Ｂへ送信するためのライト要求を発行する。ペア番号１（ペア＃１）のペア情報は，正ボリュームのある記憶制御装置２０Ａ内のペア情報テーブル２２０に登録し，ペア番号２（ペア＃２）のペア情報は，記憶制御装置２０Ｂ内のペアであるため，記憶制御装置２０Ｂ内のペア情報テーブル２２０に登録する。

【００３３】

レプリケーションを作成する場合，同一記憶制御装置内で作成することが望ましく，なるべく同一記憶制御装置内で作成するようにするが，レプリケーション元のボリュームがある記憶制御装置に集中したりした場合，同一記憶制御装置内に空き領域がなくなり，複数ある記憶制御装置間でレプリケーションを作成する場合が発生する。このような場合に異記憶制御装置間レプリケーションを作成する一実施例が図７に示すものであり，ここでは，正ボリューム３１１のレプリケーションを実副ボリューム３１２にとっている。この場合でも，正ボリューム３１１のレプリケーションを仮想副ボリューム３１３にとるようにレプリケーション先を選択することにより，同一記憶制御装置内で作成しているように見せかける。仮想副ボリューム３１３は，ディスク装置内のデータ領域は使わず，ボリューム番号だけ使っている。これが仮想的に割り付ける処理である。ホスト計算機は，正ボリューム３１１のレプリケーションを仮想副ボリューム３１３にとるように認識しているため，レプリケーション先がどの記憶制御装置かを認識せずに従来どおりに処理できる。ホスト計算機は，正ボリューム３１１と仮想副ボリューム３１３のある記憶制御装置に要求を出せばよい。一方，記憶装置システム７０側では仮想副ボリューム３１３は仮想的に割り付けられたボリュームであることと，実際のボリュームは実副ボリューム３１２であることを認識している。このため，実副ボリューム３１２への要求がきたら，仮想副ボリューム３１３の記憶制御装置で受け，実副ボリューム３１２のある記憶制御装置へ要求内容を送信することになる。

【００３４】

図８は，異記憶制御装置間レプリケーション作成処理方式を示すフローチャート例である。前記ステップ５０４０の異記憶制御装置間レプリケーション作成処理方式を図８のフローチャートに従って説明する。

【００３５】

ステップ８０１０～８０７０は，ステップ６０１０～６０７０と同じである。図８には詳細なステップは図示していないが，次のように処理が行われる。ステップ８０１０（ステップ６０１０に対応）で図２，図１１に示す差分ビットマップ２４０Ｐ１の全ビットを‘1’とし，ステップ８０２０（ステップ６０２０に対応）で差分ビットマップ上の‘1’を検出したら，ステップ８０３０（ステップ６０３０に対応）でそのビットに対応する箇所のデータがキャッシュにあるか否かの判定を行い，ステップ８０２０で‘1’を検出しなければ，ステップ８０７０（ステップ６０７０に対応）へ進む，ステップ８０３０でキャッシュになければステップ８０４０（ステップ６０４０に対応）で正ボリュームからキャッシュにリードし，ステップ８０５０（ステップ６０５０に対応）で仮想化したキャ

ッシュ内で副ボリュームのデータ用にコピーし，キャッシュに格納したら，ステップ８０６０（ステップ６０６０に対応）で差分ビットを‘0’にし，ステップ８０７０（ステップ６０７０に対応）で，次のビットがあれば，ステップ８０２０から８０６０を繰り返し，次のビットが無ければ，終了する。このようにして，キャッシュ上に仮想化した副ボリューム用データが格納される。ここまでの処理は，正副ボリュームが同じ記憶制御装置内にある場合と同様である。

【００３６】

一方，非同期でキャッシュ上の仮想化した副ボリューム用データを実際の副ボリュームに格納する必要がある。そのために，正側記憶制御装置２０Ａからレプリケーションを作成するための要求を副側記憶制御装置２０Ｂへ送信し，副側記憶制御装置２０Ｂ内にレプリケーションを作成する。具体的には，次のようにして，副側記憶制御装置２０Ｂに接続されたディスク装置にレプリケーションを作成する。正ボリュームのある記憶制御装置２０Ａは，副ボリュームを作成する記憶制御装置２０Ｂの記憶装置に反映されていないキャッシュ上のデータ（ダーティデータ）をライトする要求を記憶制御装置２０Ｂへ発行する（ステップ８１１０）。副側記憶制御装置２０Ｂでは，ライトデータ格納用にキャッシュを確保する（ステップ８１２０）。キャッシュが確保できたことを正側記憶制御装置２０Ａへ報告する（ステップ８１３０）。正側記憶制御装置２０Ａは報告を受けたらライトデータを転送する（ステップ８１４０）。正側記憶制御装置２０Ａは副側記憶制御装置２０Ｂから転送完了報告を受け取る（ステップ８１５０）。正側記憶制御装置２０Ａは次のダーティデータがあれば，ステップ８１１０～ステップ８１５０を繰り返す（ステップ８１６０）。次のダーティデータが無ければ，本処理は終了する。副側記憶制御装置２０Ｂは，非同期にキャッシュからデータを副ボリュームへ格納する（ステップ８１７０）。

【００３７】

初期コピー処理は一度全ての差分ビットに対応するデータをコピーする処理である。初期コピーが終了しても，その間ライト要求を受領したら差分ビットマップに‘1’が立つ。これらの残コピー処理は，差分ビットマップを先頭から順に検索し，ビットに‘1’が立っているところを検出したら，初期コピー処理と同様のコピー処理を行う。

【００３８】

前記初期コピー中に通常のリード／ライト要求が来る可能性もある。前記ライト要求が来た場合の処理について，図９のフローチャートに従って説明する。

ホスト１０からのライト要求を記憶制御装置２０Ａが受け取る（ステップ９０１０）。記憶制御装置２０Ａは，ライト対象データに対応する箇所の差分ビットマップ２４０Ｐ１（図２，図１１参照）の対象箇所のビットを‘1’にする(ステップ９０２０)。記憶制御装置２０Ａはライトデータ格納用のキャッシュメモリ領域を確保する（ステップ９０３０）。記憶制御装置２０Ａは，ホスト１０からのライトデータを受領してキャッシュに格納する（ステップ９０４０）。記憶制御装置２０Ａは，ホスト１０へライト完了報告を返す（ステップ９０５０）。

憶制御装置２０Ａは，ホストからのライト要求とは非同期に正ボリュームへ格納する（ステップ９０６０）。

【００３９】

副ボリュームへ前記ライトデータの内容を反映する処理は，差分ビットマップを順次見てビット‘1’が検出された場合（ステップ９０７０）に，記憶制御装置内ペアであれば（ステップ９０８０），ステップ６０３０～６０６０，６０８０の処理（ステップ９０９０），異記憶制御装置間のペアではステップ８０３０～８０６０，８１１０～８１５０，８１７０の処理を行う（ステップ９１００）。次のビットを検索する場合は，ステップ９０７０へ進み，なければ終了する（ステップ９１１０）。

【００４０】

次にＳｐｌｉｔ状態の場合について説明する。Ｓｐｌｉｔ状態になったら，２４０Ｐ１とＰ２を切替えて使用する。図１１に示すように，差分ビットマップは正ボリューム側の記憶制御装置内のメモリ２２にもち，１ペアにつきＰ１とＰ２を持つ。Ｐ１とＰ２はサイ

ズが同じものとする。Ｓｐｌｉｔ状態中に使用する差分ビットマップは正ボリューム側の記憶制御装置２０Ａ内のメモリ２２にもつので，副ボリュームへの更新があった場合は，毎回記憶制御装置２０Ａ内の差分ビットマップをアクセスする。

【００４１】

正ボリュームの内容に再同期するリシンク処理（Ｒｅｓｙｎｃ）の場合，その時点の正ボリュームの内容を副ボリュームにコピーする処理であるため，正副２つの差分ビットマップをマージして初期コピーと同様の処理を行う。すなわち，正副どちらか一方のビットに‘１’が立っている場合は，全てコピー処理を行う。

【００４２】

通常Ｓｐｌｉｔは，初期コピーが終わって正／副ボリューム内容の同期がとれた時点で実行する。これに対し，初期コピー中にＳｐｌｉｔ要求を受け取った場合でも，ホスト１０へは即時Ｓｐｌｉｔ完了報告をだし，バックグラウンドで残りのコピーを実行する『高速Ｓｐｌｉｔ』がある。図１２は，本方式の場合のＳｐｌｉｔ処理を示すフローチャートである。図１２に示すように，ホスト１０が『高速Ｓｐｌｉｔ』を発行し，記憶制御装置２０Ａが受領したら（ステップ１２０１０），ホスト１０からのライト要求箇所を格納するための差分ビットマップ２４０をＰ１からＰ２へ記憶制御装置２０Ａが切り替える（ステップ１２０２０）。記憶制御装置２０Ａは，ペア情報テーブル２２０のペア状態２２６（図４参照）を‘Ｓｐｌｉｔ’に変更する（ステップ１２０３０）。これで，正／副ボリュームはリード／ライト要求を受け付け可能となる。記憶制御装置２０は，バックグラウンドで副ボリュームへ未反映のデータをコピーする処理を実行する。副ボリュームへ反映する処理は，差分ビットマップを順次見てビット‘１’が検出された場合に（ステップ１２０４０），同一記憶制御装置２０内のペアであれば（ステップ１２０５０），ステップ６０２０～６０８０と同様の処理を行う（ステップ１２０６０）。異記憶制御装置２０間のペアの場合，ステップ８０２０～８０７０及びステップ８１１０～８１６０の処理を実行する（ステップ１２０７０）。次のビットを検索する場合は，ステップ１２０４０へ進み，なければ処理を終了する（ステップ１２０８０）。

【００４３】

前記高速Ｓｐｌｉｔ状態になったときにホスト１０から正ボリュームがライト要求を受け取ったときの処理を図１３に示す。

ホスト１０から記憶制御装置２０Ａは正ボリュームへのライト要求を受領する（ステップ１３０１０）。記憶制御装置２０Ａは，差分ビットマップ２４０Ｐ１（図１１参照）のライト対象データに対応する箇所のビットを‘１’に設定する（ステップ１３０２０）。そのビットに対応する箇所のデータがキャッシュにあるか否かの判定を行う（ステップ１３０３０）。ステップ１３０３０でキャッシュになければ正ボリュームからキャッシュにリードする（ステップ１３０４０）。差分ビットマップＰ２のライト対象旧データに対応する箇所のビットが‘１’を検出したらステップ１３０６０へ進み，検出しなければステップ１３０８０へ進む（ステップ１３０５０）。‘１’が立っている場合は，そのビットに対応するデータがまだ副ボリュームへコピーし終わっていないことを示す。キャッシュ内で副ボリュームのデータ用領域（異記憶制御装置間のペアの場合は，キャッシュ内で仮想化した副ボリューム用領域）に旧データをコピーする（ステップ１３０６０）。または，キャッシュにデータがなければ，正副用に２回リードしてもよい。コピーに伴い，データが正しいか否かを判定するための冗長情報も副ボリューム用に新規に作成しデータへ付随させる。キャッシュに格納したら，差分ビットを‘０’にする（ステップ１３０７０）。以上のステップ１３０３０～１３０７０の処理により，正ボリュームの旧データがキャッシュ内の副ボリューム用データ領域（異記憶制御装置間のペアの場合は，キャッシュ内の仮想化した副ボリューム用データ領域）に格納されたことになる。ステップ１３０７０の後，または，ステップ１３０５０の判断でＮｏの場合，対象となったビットへのライトデータをホスト１０から受け取り，正ボリューム用のキャッシュ領域に格納する（ステップ１３０８０）。ホスト１０へライト完了報告を出す（ステップ１３０９０）。

【００４４】

同一記憶制御装置内のペアであれば（ステップ１３１００），非同期でキャッシュ上のデータを副ボリュームに格納する（ステップ１３１１０）。異記憶制御装置間のペアの場合，実際のデータが格納されている副ボリュームのある記憶制御装置２０Ｂに，前記キャッシュ上の旧データを格納するため，ステップ８１１０～８１６０の処理を実行する（ステップ１３１２０）。

【００４５】

次に，前記高速Ｓｐｌｉｔ状態になったときにホスト１０から副ボリュームがライト要求を受け取ったときの処理を説明する。フローチャートは図１３と同じである。

ステップ１３０１０と同様に，ホスト１０から記憶制御装置２０Ａは仮想化副ボリュームへのライト要求を受領する。ステップ１３０２０と同様に，記憶制御装置２０Ａは，差分ビットマップ２４０Ｐ１（図１１参照）のライト対象データに対応する箇所のビットを‘1’に設定する。ステップ１３０３０と同様に，ビットに対応する箇所のデータが正ボリューム用キャッシュ領域にあるか否かの判定を行う。ステップ１３０４０と同様に，ステップ１３０３０でキャッシュになければ正ボリュームからキャッシュにリードする。ステップ１３０５０と同様に，差分ビットマップＰ２のライト対象旧データに対応する箇所のビットが‘1’を検出したらステップ１３０６０へ進み，検出しなければステップ１３０８０へ進む。ステップ１３０６０と同様に，キャッシュ内で副ボリュームのデータ用領域（異記憶制御装置間のペアの場合は，キャッシュ内で仮想化した副ボリューム用領域）に旧データをコピーする。または，キャッシュにデータがなければ，正副用に２回リードしてもよい。または，キャッシュからリードする場合，直接キャッシュ内で副ボリュームのデータ用領域（異記憶制御装置間のペアの場合は，キャッシュ内で仮想化した副ボリューム用領域）に旧データをリードしてもよい。そうすることで，リード処理が１回ですむ。コピーに伴い，データが正しいか否かを判定するための冗長情報も副ボリューム用に新規に作成しデータへ付随させてキャッシュに格納する。もし，ライトデータがビットに対応する全データと一致するならば，キャッシュに旧データがなくてもよいため，領域だけ確保する。キャッシュに格納したら，差分ビットを‘0’にする（ステップ１３０７０と同様）。対象となったビットへのライトデータを受け取り，副ボリューム用のキャッシュ領域に格納する（ステップ１３０８０と同様）。ホスト１０へライト完了報告を出す（ステップ１３０９０と同様）。後はステップ１３１００～１３１２０と同様である。副ボリュームへのライト要求を出すホストは正ボリュームへライト要求を出すホスト１０と別のホストでもよい。

【００４６】

次に，ボリュームの再同期について図１５のフローチャートに従って説明する。Ｓｐｌｉｔ状態になった正ボリュームと副ボリュームはそれぞれリード／ライト要求を処理しているため，内容が異なるボリュームとなっている。再同期とは，その時点の正ボリュームの内容に副ボリュームの内容を合わせることを示す。最初にペア情報のペア状態を変更する（ステップ１５０１０）。差分ビットマップ２４０Ｐ１とＰ２をマージして，差分ビットマップ２４０Ｐ１に格納する（ステップ１５０２０）。差分ビットマップ２４０Ｐ１の‘1’を検出したら（ステップ１５０３０），初期コピー同様の処理を行う。具体的には，同一記憶制御装置内ペアであれば（ステップ１５０４０），ステップ６０２０～６０８０（ステップ１５０５０），異記憶制御装置間ペアの場合はステップ８０２０～８１６０となる（ステップ１５０６０）。次のビットを検索する場合は，ステップ１５０３０へ進み，なければ本処理を終了する（ステップ１５０７０）。

【００４７】

以上説明したように，コピー元とコピー先のクラスタが異なる場合，実際のコピー先ボリュームをコピー元のクラスタ内にある空きボリュームに仮想的に割り付ける仮想化技術により，同一クラスタ内のペアとしてホストに見せることができる。それにより，クラスタを意識することなく自由に記憶領域のコピーを作成するクラスタ構成記憶装置システムを提供できる。

【００４８】

上記にレプリケーション作成処理を説明したが，以下，処理速度優先処理について説明する。処理速度優先処理は，コピーボリュームを作成する処理と通常リード／ライト要求（通常Ｉ／Ｏ）を，状況に応じた優先順位に従って処理できるようにするものである。

【００４９】

前記レプリケーション処理により，コピーに伴う処理を副クラスタの状況にかかわらず，正クラスタから副クラスタに出してしまうが，副クラスタは，通常のホストＩ／Ｏを受付処理している。また，副クラスタ内でもコピー機能のエンジンがあり，当該クラスタ内のボリュームのコピーをクラスタ内で作成している。このように，クラスタ毎にコピー機能をもち，ジョブへの要求キューを持ち，ジョブ種類の優先順位に従ってジョブを起動し処理している。

【００５０】

クラスタ間にまたがったコピー処理を行う場合，クラスタ間またがりで通信する形態を通常のリード／ライト要求形式で行うと，要求を受け取った側は，ホストからの通常Ｉ／Ｏと区別できない。しかし，コピー機能系の処理よりホストＩ／Ｏを優先させて実行したい。

【００５１】

すなわち，異記憶制御装置間のペアの場合，副ボリュームのある記憶制御装置２０Ｂは，正側記憶制御装置２０Ａから受け取ったライト要求とホスト１０から受け取った通常ライト要求の区別がつかない。従って，記憶制御装置２０Ｂは，全てのライト要求は受け取った順番に処理する。これにより，通常ライト要求の応答時間が遅くなり，性能が低下したように見えてしまうことがある。これらを記憶制御装置２０Ｂが区別して実行順序をスケジュールすることにより，高性能化を実現する。

【００５２】

以下に示す実施例は，処理優先度をコピー元がコピー先に設定する処理速度優先処理に関するものである。

正側記憶制御装置２０Ａから受け取ったライト要求とホスト１０から受け取った通常ライト要求を区別する単位として，ボリューム単位を考える。記憶制御装置２０Ｂ内にあるボリュームでレプリケーション作成中のボリュームの場合，記憶制御装置２０Ａの側で，ペアを組んでいるかどうかを認識している。記憶制御装置２０Ｂのレプリケーション作成中のボリュームへの要求は，レプリケーションするためのコピー処理なのか，通常リード／ライトなのか，ペア状態により記憶制御装置２０Ａ内で判断できる。

【００５３】

これを図１６のフローチャートに従って説明する。記憶制御装置２０Ａ内の要求処理優先度設定プログラム２０２（図２参照）により，前記２０Ａ内のボリュームとレプリケーションを組んでいる記憶制御装置２０Ｂ内のボリュームへの要求を通常リード／ライトと同等に処理すべきか否かの処理優先度を判断する（ステップ１６０１０）。記憶制御装置２０Ａ内で判断した処理優先度情報を記憶制御装置２０Ｂへ渡す（ステップ１６０２０）。記憶制御装置２０Ｂは，メモリ２２内にある処理優先度情報ビットマップ２５０（図２，図１４参照）に前記処理優先度情報を対象としているボリューム番号のビット位置に登録する（ステップ１６０３０）。処理優先度情報ビットマップ２５０は，ボリューム番号毎の，通常リード／ライトと同様にスケジュールして処理するか否かの情報を管理する。例えば，急いで処理したいので，通常リード／ライトと区別なくスケジュールする場合はビット‘１’，通常リード／ライトを優先させたい場合は，‘０’とする。また，通常リード／ライトを優先度５として，それ以外に優先度４～１で管理してもよい。その際は，１ボリュームにつき１ビットではなく，数ビット用意する。

【００５４】

レプリケーション作成中である場合，ステップ８１１０で，記憶制御装置２０Ａがリード／ライト要求を記憶制御装置２０Ｂに発行すると（ステップ１６０４０），記憶制御装置２０Ｂは前記要求を受け取り，要求の対象となるボリュームの処理優先度情報を参照する（ステップ１６０５０），処理優先度情報が通常リード／ライト要求と同等にスケジュ

ール可能であれば（ステップ１６０６０），通常リード／ライト処理キュー２６１（図２参照）に入れ（ステップ１６０７０），キューから要求が選択されたら，ステップ８１２０のキャッシュエリアの確保を行い，ステップ８１３０～８１５０を実行する（ステップ１６０８０）。ステップ１６０６０で処理優先度情報が通常リード／ライト要求より低ければ，通常リード／ライト処理キュー２６１と別の処理優先度低キュー２６２（図２参照）に入れ（ステップ１６０９０），選択されたら，ステップ１６０８０の処理を行う。ボリュームの処理優先度情報度はＳＶＰ（SerVice Processor）などの端末，例として管理端末８０から保守員により入力してもよい。

通常リード／ライト処理キュー２６１及び処理優先度低キュー２６２は，配列またはリスト構造により，発行時刻順に管理されるキュー構造の制御情報である。

【００５５】

仮想化したボリューム３１３（図７参照）に対応する実の副ボリューム３１２は，ホスト１０からは見えていないボリュームになるため，ホスト１０からボリューム３１２に対して要求が発行されることはない。ボリューム３１２への要求は必ず記憶制御装置２０Ａを介してしかこない。従って，ボリューム３１２の要求の処理を急ぐべきか否かを記憶制御装置２０Ａ側で判断できる。

【００５６】

前記処理優先度低キューのスケジュールについて図１７のフローチャートに従って説明する。処理優先度低キュー２６２は通常リード／ライト処理キュー２６１より，処理頻度を低くする。処理頻度αはユーザの設定によって変えられる。

【００５７】

前記処理頻度をαとする。通常リード／ライト処理キュー２６１に未処理な要求があるか判定する（ステップ１７０１０）。あれば，カウンタｃ＝０を設定する（ステップ１７０２０）。α＞ｃであれば（ステップ１７０３０），通常リード／ライト処理キュー２６１の処理を実行する（ステップ１７０４０）。カウンタｃをインクリメントする（ステップ１７０５０）。通常リード／ライト処理キュー２６１に未処理な要求があるか判定する（ステップ１７０６０）。あれば，ステップ１７０３０へ進む。無ければ，ステップ１７１００へ進む。ステップ１７０３０でα≦ｃとなったら，処理優先度低キュー２６２の未処理キュー数≧ｎ，又は，処理優先度低キュー２６２の時刻の古い未処理キューがｍ時間以上過ぎているか判定する（ステップ１７０７０）。ｍ，ｎは予め決めた数である。ステップ１７０７０で条件に当てはまれば，処理優先度低キュー２６２の処理を実行する（ステップ１７０８０）。カウンタｃを０にセットし，ステップ１７０６０へ進む（ステップ１７０９０）。ステップ１７０７０で条件に当てはまらなければステップ１７０９０へ進む。

ステップ１７０１０で未処理な要求がないと判定された場合，処理優先度低キュー２６２の処理を実行し，ステップ１７０１０へ進む（ステップ１７１１０）。

【００５８】

また，ステップ１６０２０で記憶制御装置２０Ａから記憶制御装置２０Ｂへリード／ライト要求を発行する場合，記憶制御装置２０Ａは，記憶制御装置２０Ｂの状態を意識せずに発行する。例えば，記憶制御装置２０Ｂが通常リード／ライト処理や記憶制御装置２０Ｂ内のレプリケーション処理によって，記憶制御装置２０Ｂのキャッシュのダーティ量が多くなっている場合がある。ダーティ量が一定基準値より多くなるということは，キャッシュの空き容量が少なくなるということで，レプリケーション処理を一時とめたりして，キャッシュからディスク装置へデータを確保する処理を優先処理してキャッシュの空き領域を増やす必要がある。記憶制御装置２０Ｂはステップ１６０５０で要求の対象となるボリュームの処理優先度を参照すると共に，ダーティ量を参照し，ダーティ量がある一定基準値（あらかじめ設定しておく）を超えていれば，ステップ８１２０～８１５０のキャッシュを確保してデータを受領する処理を遅らせることにより，記憶制御装置２０Ａからの要求の受け入れ量を調整する。記憶制御装置２０Ａは記憶制御装置２０Ｂから完了報告を受取るまでは次の要求は出さない。

【００５９】
次に，正クラスタが副クラスタ側にコピー先ボリュームへの処理が通常リード／ライト要求（通常Ｉ／Ｏ）に影響がないように処理するか，通常リード／ライト要求（通常Ｉ／Ｏ）と同等に処理するかの情報を与える処理速度優先処理の変形例を示す。前述の実施例とは異なり，以下の変形例はコピー先が処理を優先すべきか判断する処理速度優先処理の例である。
【００６０】
正クラスタ側から副クラスタ側に送信する要求がコピー処理要求（レプリケーションを作成するための要求）である場合は，正クラスタ側（コピー元）が要求命令に本要求がコピー処理要求（レプリケーションを作成するための要求）であることを示す要求種別を付加し副クラスタ側に送信する。これを受けた副クラスタ側（コピー先）が，副クラスタ側は要求種別が付加されている要求は優先度が低い要求と認識し，処理優先度低キュー２６２（図２参照）に入れる。これにより，正クラスタ側から受信した要求種別に基づいて，副クラスタ側が処理優先度を決めることができる。この例では，正クラスタ側は，レプリケーションを作成するための要求にその要求がレプリケーションを作成するための要求であることを示す情報を付加して，その要求を送信するが，優先度が低いかどうかを判断するのはあくまで副クラスタ側である。したがって，副クラスタ側でレプリケーションを作成するための要求に対して優先度を低く処理するかどうかを決定できる。
【００６１】
他の方法としては，正クラスタ側（コピー元）が，コピー処理要求であるという種別ではなく，汎用的に，要求の優先／非優先の情報または優先度ｎを命令に付加し，これを受けた副クラスタ側（コピー先）で前記付加された優先度情報を認識して，その優先度を考慮した処理順序のスケジュールを行う。これにより，副クラスタ側が，正クラスタ側から受信した要求の優先／非優先の情報または優先度ｎに基づいて，処理優先度を決めることができる。この例の場合も，レプリケーションを作成するための要求に対して優先度を低く処理するかどうかを決定するのは副クラスタ側である。しかし，この例の場合は，一般的な要求に含まれる要求の優先／非優先の情報または優先度ｎに基づいて副クラスタ側が判断するので，副クラスタ側で通常の要求の優先処理によりレプリケーションを作成するための要求を優先処理するかどうかを決定できる。
【００６２】
正クラスタが，副クラスタ側にコピー先ボリュームへの処理が通常Ｉ／Ｏの影響が無い様に処理するか，通常Ｉ／Ｏと同等に処理するかの情報を与え，副クラスタ側で前記情報を参照して，前者の場合は，コピーに伴う処理は優先度を落として処理し，後者の場合は発生時刻順に処理するようなスケジュールする。
【００６３】
以上説明した処理速度優先処理により，通常業務のＩ／Ｏを妨げないでレプリケーションを行うことができる。
【００６４】
以上，本発明の実施例のクラスタ構成記憶装置システムについて説明したが，以下，前記クラスタ構成記憶装置システムの仮想化技術により，クラスタ間のペアを同一クラスタ内のペアとして実現する際の，ユーザの使い方を図１８により説明する。
【００６５】
最初に，ユーザはコピー先ボリュームを選択する。コピー元ボリューム３１１に対し，コピー先ボリューム候補である空きボリューム群３２１，３２２，３２３，３２４は副ボリュームプールとしてユーザに提示される。副ボリュームプールはコピー元ボリュームと異なるクラスタのボリュームも含む。
【００６６】
ユーザがプールより副ボリュームを選択する。例えば３２２を管理端末８０から入力することにより選択する。図１８の例では，選択された３２２はコピー元ボリューム３１１が属するクラスタ（記憶制御装置２０Ａ）とは異なったクラスタ（記憶制御装置２０Ｂ）

に属する。

【００６７】

コピー先ボリュームが異なるクラスタのボリュームとなるため，仮想化する副ボリュームをコピー元のクラスタ内の空きボリュームから選択する。ここではボリューム３１２を仮想副ボリュームとして選択する。このボリュームは実体を持たないため，ＨＤＤなどの物理リソースは必要がない。従って，仮想化用の専用のデバイス番号を予め用意し，そこから選択するようにしてもよい。

【００６８】

また，この例ではコピー元と同じクラスタに副ボリューム候補があるので，まずそのボリュームから選択できるように，ユーザに提示するとき，候補の中からシステム側で選択して提示してもよい。また，ユーザへのボリューム候補提示や副ボリュームプールからの選択または仮想ボリュームの選択またはその両方を，管理端末８０から入力して行う方法やホストコマンドによって行う方法や記憶制御装置側で自動的に行う方法がある。

【００６９】

レプリケーション作成プログラム２０１（図２参照）は，前述のように管理端末から入力された情報，ホストからのホストコマンドに基づいてペア情報テーブル２２０への情報の登録を行う。あるいは，レプリケーション作成プログラム２０１は，自動的にペア情報テーブル２２０への情報の登録を行う。

【００７０】

以上，本発明者によってなされた発明を，前記実施例に基づき具体的に説明したが，本発明は，前記実施例に限定されるものではなく，その要旨を逸脱しない範囲において種々変更可能であることは勿論である。

【図面の簡単な説明】

【００７１】

【図１】本発明の一実施例を示す計算機システムの構成図である。

【図２】メモリの一実施例の構成図である。

【図３】管理端末（例えば，管理サーバ）の一実施例の構成図である。

【図４】ペア情報テーブルの一例を示す図である。

【図５】レプリケーション作成処理の概要を示すフローチャートの例である。

【図６】同一記憶制御装置内でのレプリケーション作成処理方式を示すフローチャートの例である。

【図７】異記憶制御装置内でのレプリケーション作成処理方式の一実施例を示す図である。

【図８】異記憶制御装置内でのレプリケーション作成処理方式を示すフローチャートである。

【図９】初期コピー中にライト要求が発行された場合の処理を示すフローチャートの例である。

【図１０】ボリューム情報テーブルの一例を示す図である。

【図１１】差分ビットマップの一例を示す図である。

【図１２】Ｓｐｌｉｔ処理を示すフローチャートの例である。

【図１３】高速Ｓｐｌｉｔ状態の正ボリュームへのライト要求受領時の処理及び副ボリュームへのライト要求受領時の処理のフローチャートの例である。

【図１４】処理優先度情報ビットマップの一例を示す図である。

【図１５】再同期処理を示すフローチャートの例である。

【図１６】処理優先度情報登録処理及び処理優先度情報によるリード／ライト処理のフローチャートの例である。

【図１７】処理優先度低キューのスケジュール方式のフローチャートの例である。

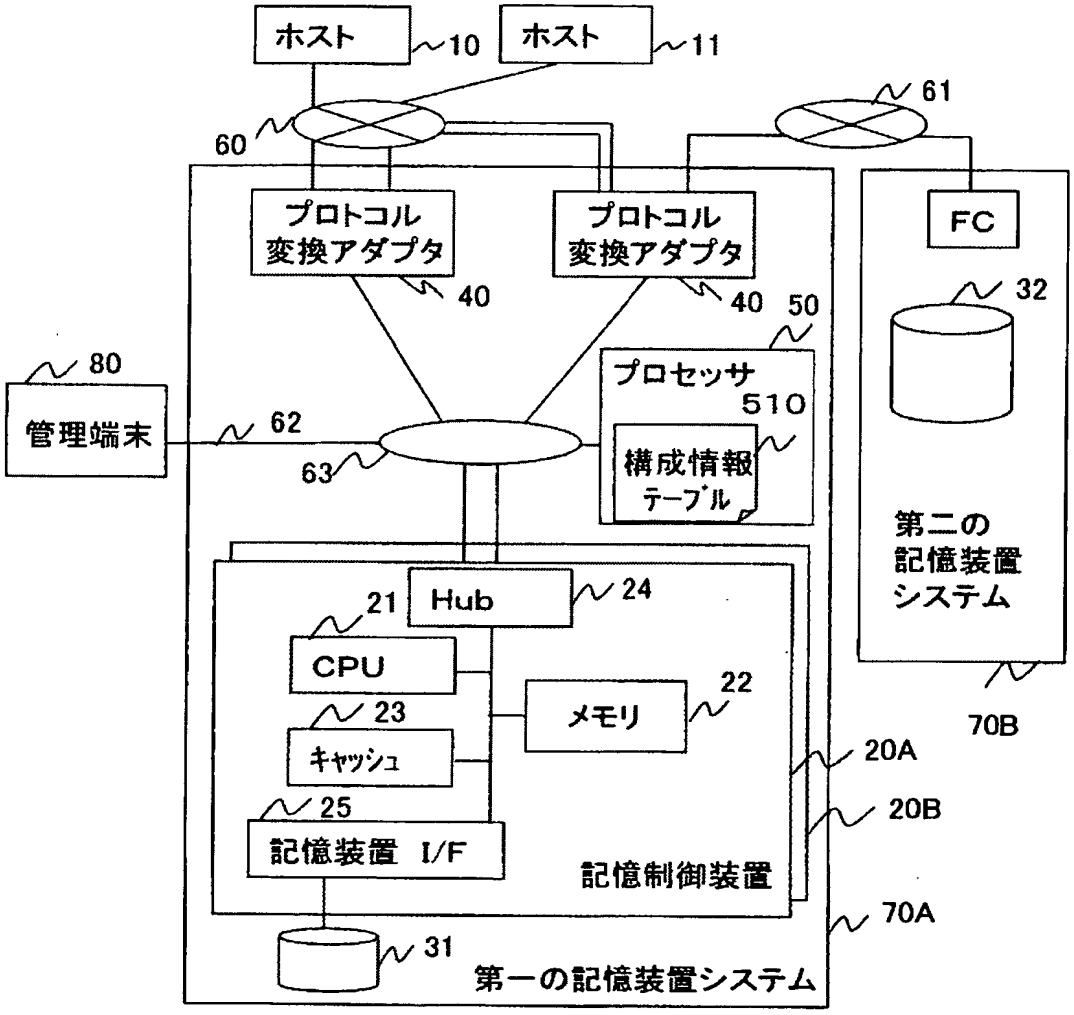
【図１８】クラスタ間のペアを同一クラスタ内のペアとして実現する際の，ユーザの使い方を説明するための図である。

【符号の説明】

【００７２】
１０…ホスト，２０…記憶制御装置，２０２…処理優先度設定プログラム，２２０…ペア情報管理テーブル，２３０…ボリューム情報テーブル，２４０…差分ビットマップ，２５０…処理優先度情報ビットマップ，３１…記憶装置，４０…プロトコル変換アダプタ，５０…プロセッサ，６０，６１，６２，６３…ネットワーク，７０…記憶装置システム，８０…管理端末

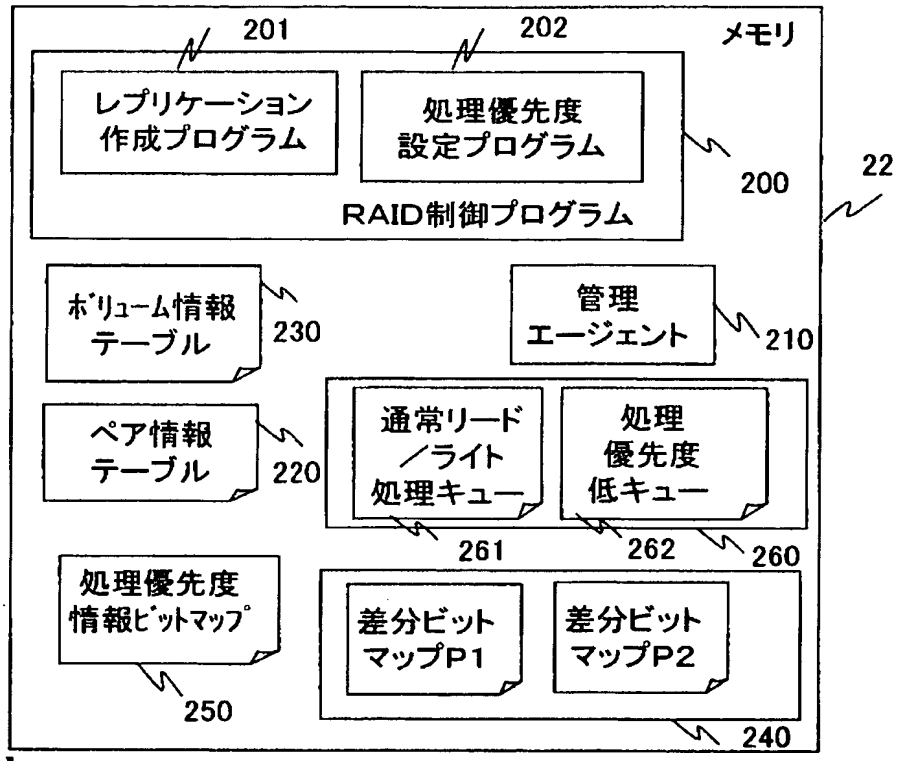
【書類名】図面
【図１】

図1


ホスト
10
ホスト
11
61
60
プロトコル
変換アダプタ
40
プロトコル
変換アダプタ
40
50
FC
32
80
管理端末
62
63
プロセッサ
510
構成情報
テーブル
第二の
記憶装置
システム
21
Hub
24
CPU
22
メモリ
23
キャッシュ
20A
20B
25
記憶装置 I/F
記憶制御装置
70A
31
第一の記憶装置システム
70B

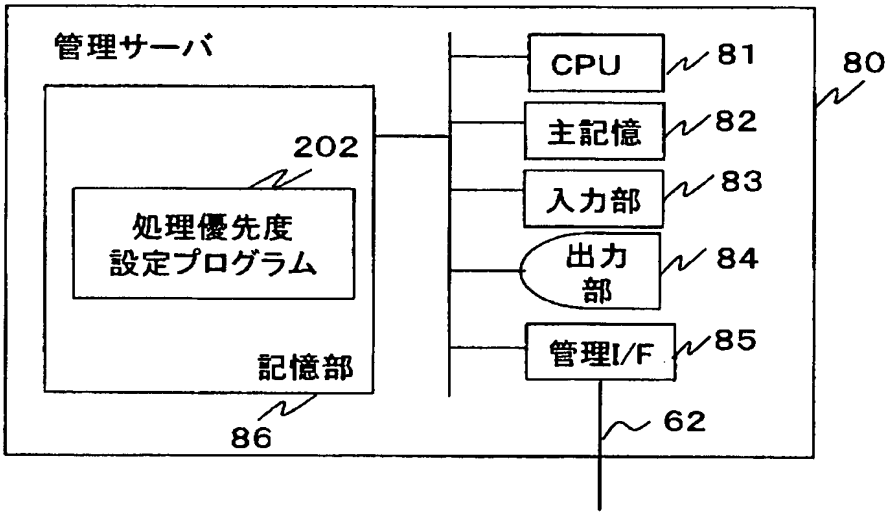
【図２】

図2


201
202
メモリ
レプリケーション
作成プログラム
処理優先度
設定プログラム
200
22
RAID制御プログラム
ボリューム情報
テーブル
230
管理
エージェント
210
ペア情報
テーブル
220
通常リード
／ライト
処理キュー
処理
優先度
低キュー
261
262
260
処理優先度
情報ビットマップ
250
差分ビット
マップP1
差分ビット
マップP2
240

【図３】

図3


管理サーバ
CPU
81
80
202
主記憶
82
処理優先度
設定プログラム
入力部
83
出力
部
84
管理I/F
85
記憶部
86
62

【図４】

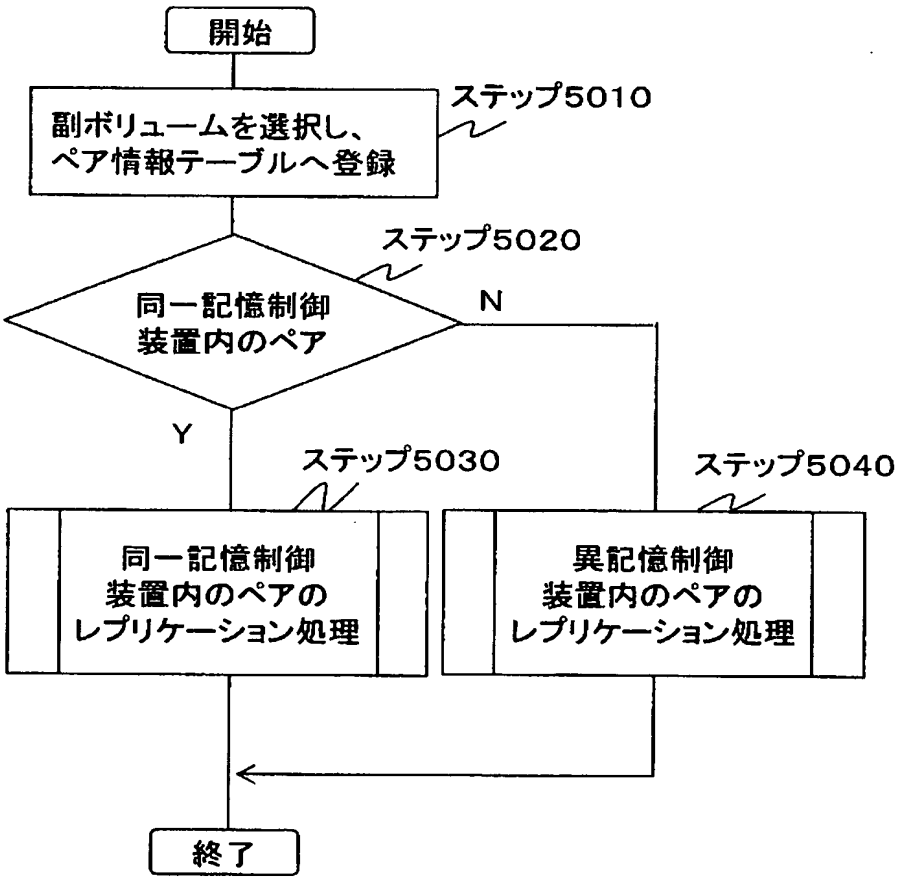
図4

220

ペア番号	正ボリューム番号	副ボリューム情報			ペア状態
		仮想化	実データ格納		
		ボリューム番号	記憶制御装置番号	ボリューム番号	
0	100	200			Pair
1	110	210	2	120	Pair
2	120	220			Split
3	130	230	3	260	Pair
4	140	240			Split
221	222	223	224	225	226

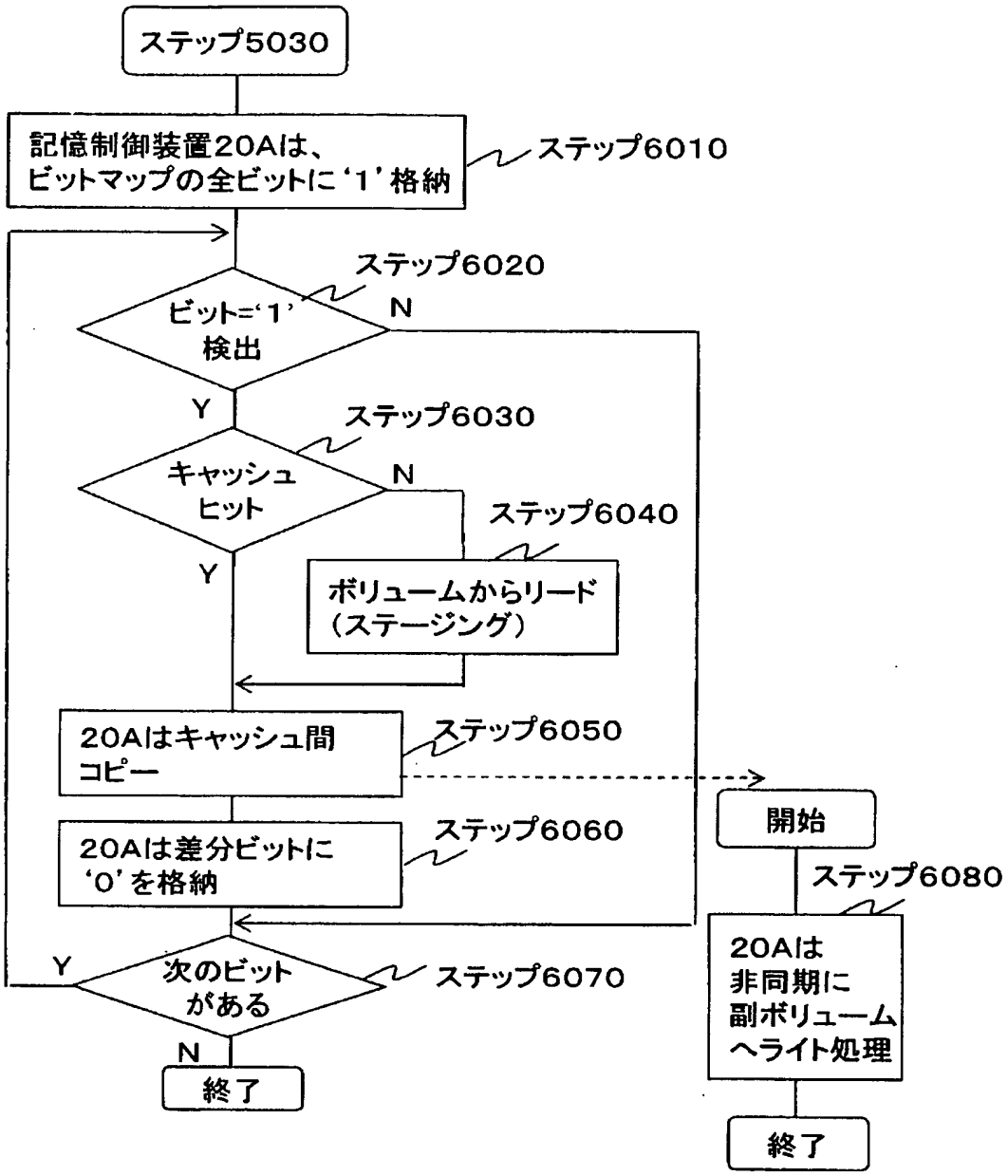
【図５】

図5



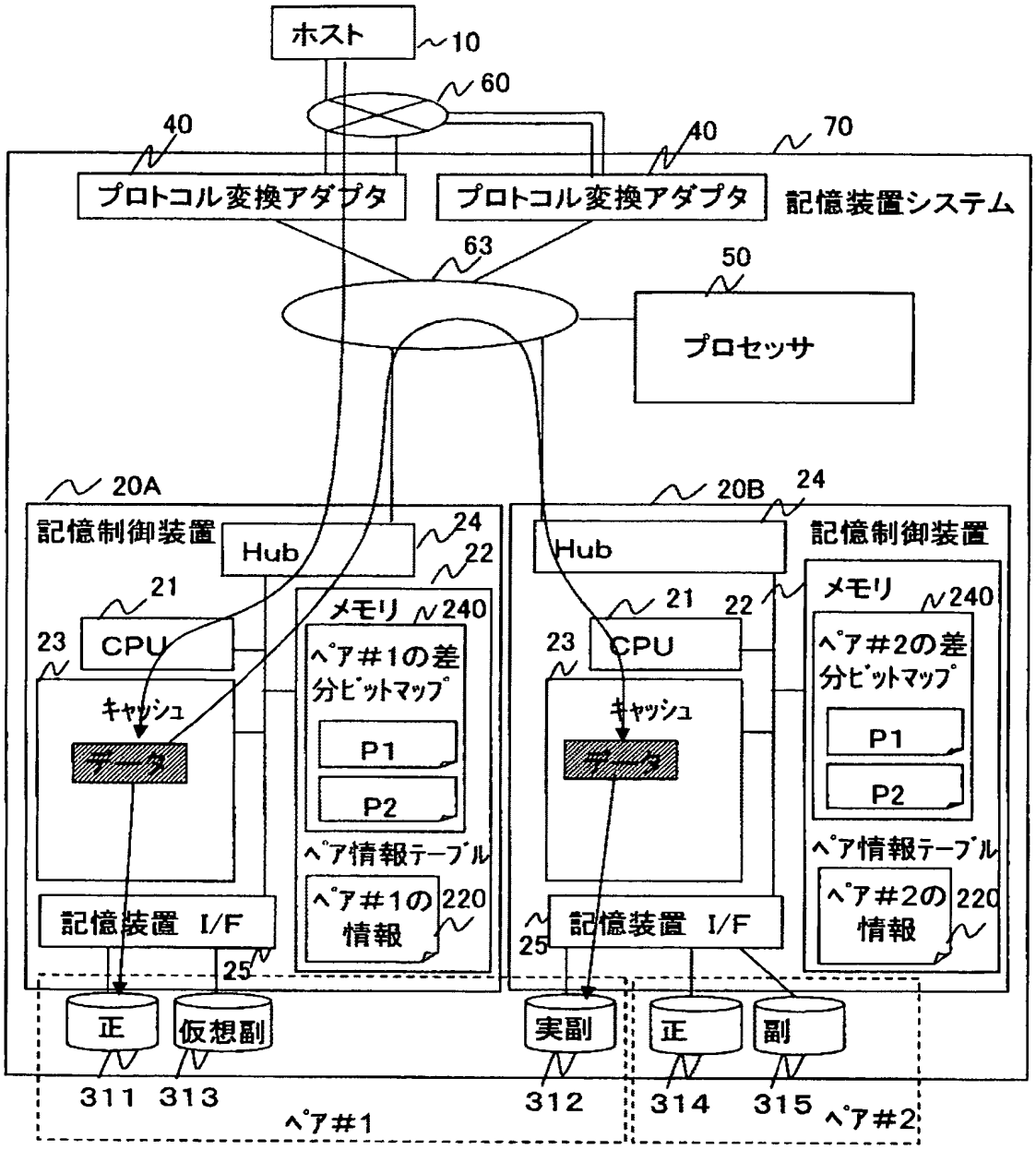
【図６】

図6


ステップ5030
記憶制御装置20Aは、
ビットマップの全ビットに‘1’格納
ステップ6010
ステップ6020
ビット=‘1’
検出
N
Y
ステップ6030
キャッシュ
ヒット
N
ステップ6040
Y
ボリュームからリード
（ステージング）
20Aはキャッシュ間
コピー
ステップ6050
20Aは差分ビットに
‘0’を格納
ステップ6060
Y
次のビット
がある
ステップ6070
N
終了
開始
ステップ6080
20Aは
非同期に
副ボリューム
へライト処理
終了

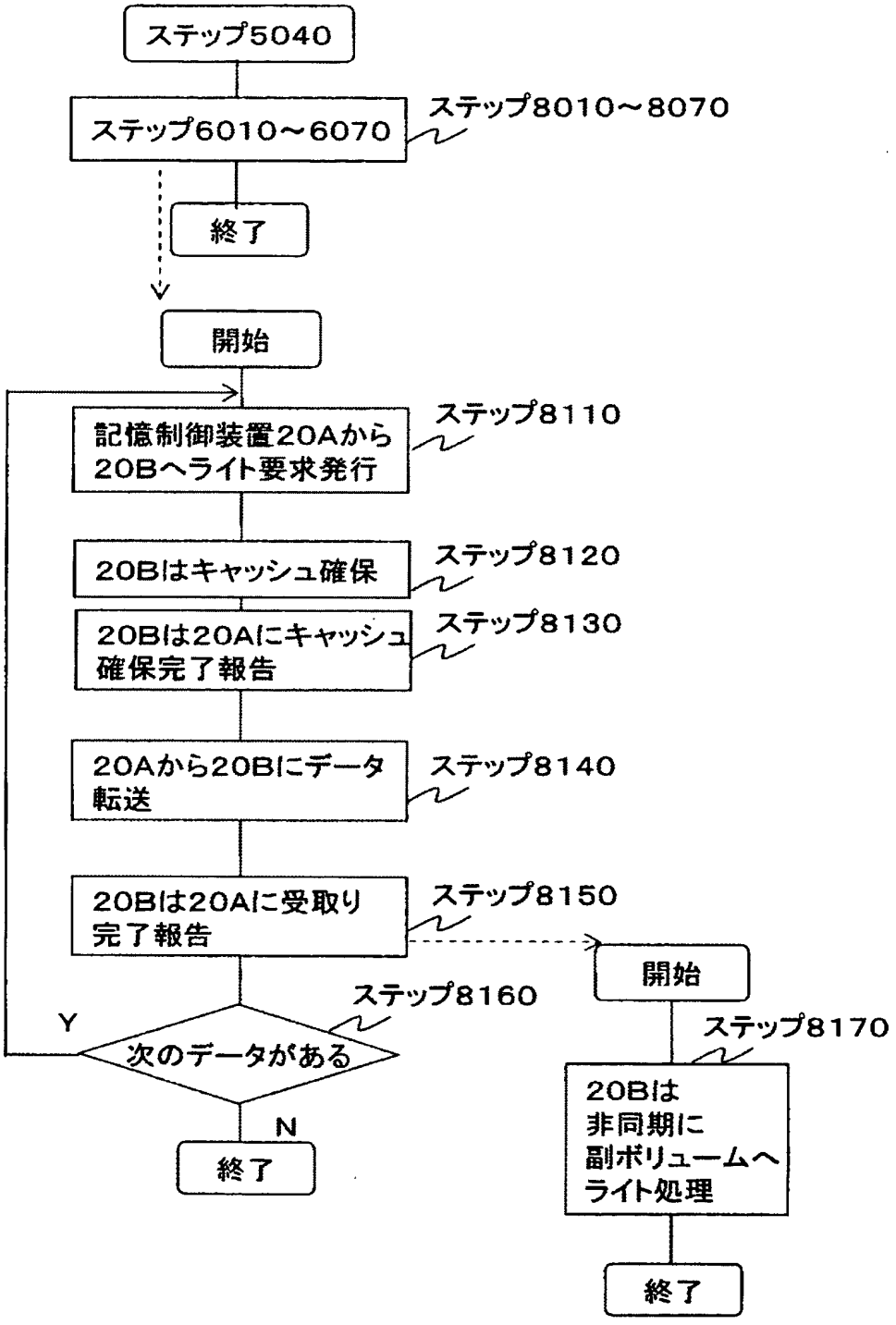
【図７】

図7


ホスト
10
60
40
40
70
プロトコル変換アダプタ
プロトコル変換アダプタ
記憶装置システム
63
50
プロセッサ
20A
20B
24
記憶制御装置
Hub
24
22
Hub
記憶制御装置
メモリ
240
21
22
メモリ
240
23
CPU
21
23
CPU
キャッシュ
データ
ペア#1の差分ビットマップ
P1
P2
キャッシュ
データ
ペア#2の差分ビットマップ
P1
P2
ペア情報テーブル
ペア#1の情報
220
ペア情報テーブル
ペア#2の情報
220
記憶装置 I/F
25
記憶装置 I/F
25
正
仮想副
実副
正
副
311
313
312
314
315
ペア#1
ペア#2

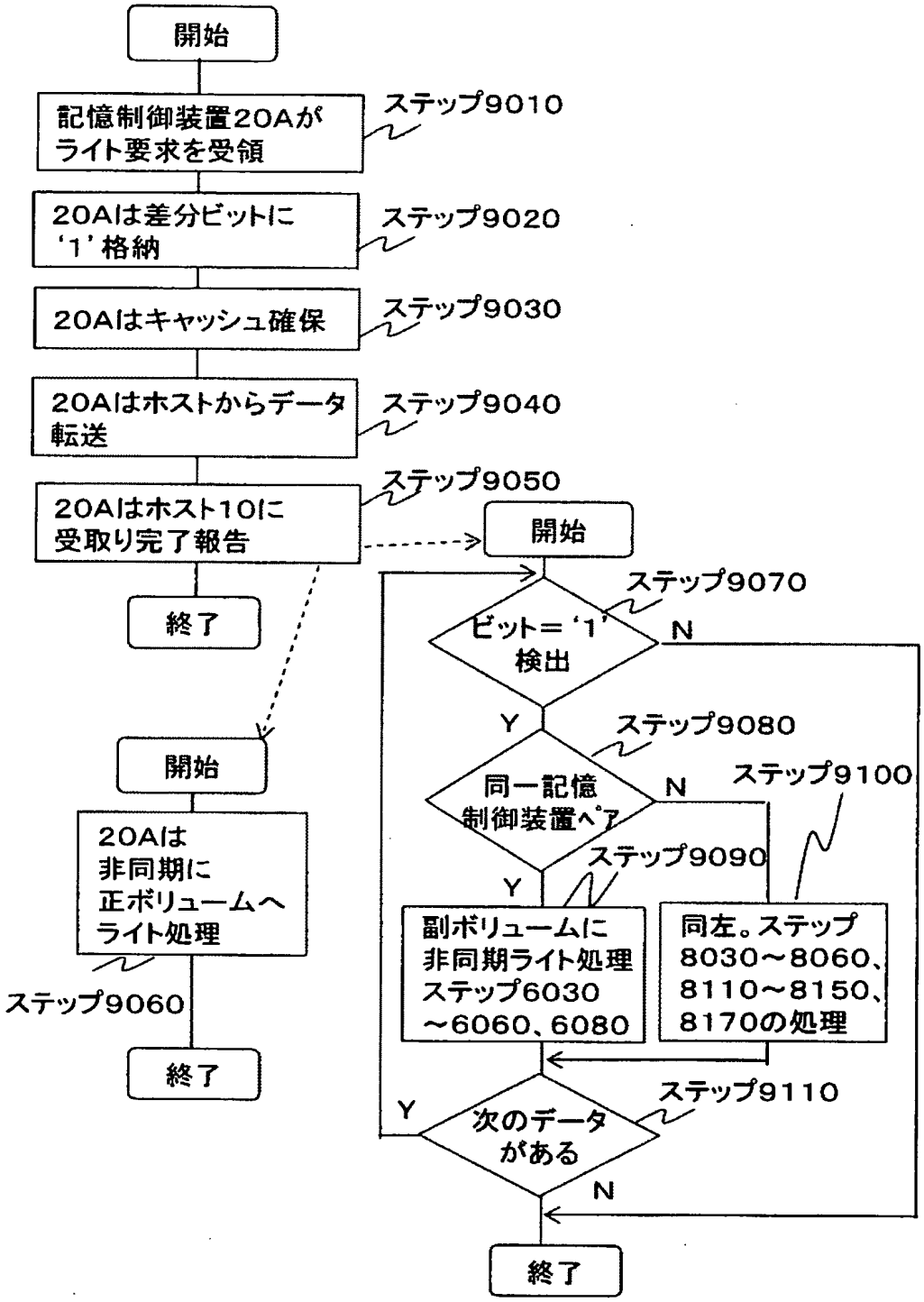
【図８】

図8


ステップ5040
ステップ6010～6070
ステップ8010～8070
終了
開始
記憶制御装置20Aから20Bへライト要求発行
ステップ8110
20Bはキャッシュ確保
ステップ8120
20Bは20Aにキャッシュ確保完了報告
ステップ8130
20Aから20Bにデータ転送
ステップ8140
20Bは20Aに受取り完了報告
ステップ8150
ステップ8160
Y
次のデータがある
N
終了
開始
ステップ8170
20Bは非同期に副ボリュームへライト処理
終了

【図９】

図9


開始
記憶制御装置20Aが
ライト要求を受領
ステップ9010
20Aは差分ビットに
'1'格納
ステップ9020
20Aはキャッシュ確保
ステップ9030
20Aはホストからデータ
転送
ステップ9040
20Aはホスト10に
受取り完了報告
ステップ9050
終了
開始
20Aは
非同期に
正ボリュームへ
ライト処理
ステップ9060
終了
開始
ステップ9070
ビット＝'1'
検出
N
Y
ステップ9080
同一記憶
制御装置ペア
N
ステップ9100
Y
ステップ9090
副ボリュームに
非同期ライト処理
ステップ6030
～6060、6080
同左。ステップ
8030～8060、
8110～8150、
8170の処理
ステップ9110
Y
次のデータ
がある
N
終了

【図１０】

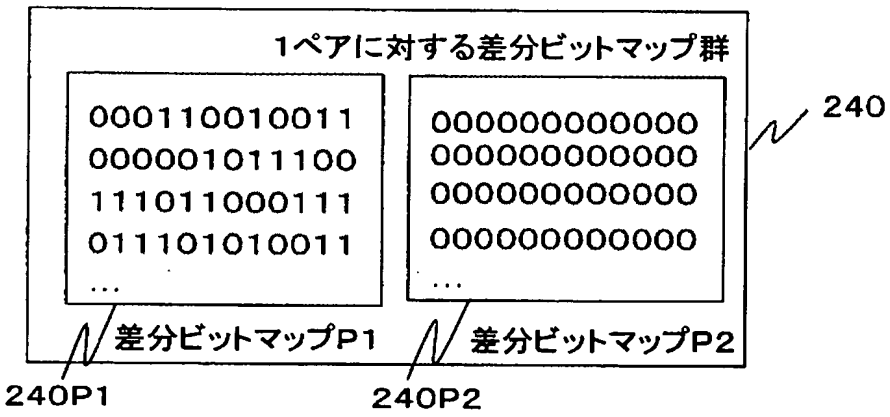
図10

ﾎﾞﾘｭｰﾑ番号	正/副	相手ﾎﾞﾘｭｰﾑ情報			ﾎﾞﾘｭｰﾑ使用中
		副ﾎﾞﾘｭｰﾑor仮想化	実ﾎﾞﾘｭｰﾑ情報		
			記憶制御装置番号	ﾎﾞﾘｭｰﾑ番号	
0	正	1024	1	20	使用中
0	正	158	—	—	使用中
0	正	1025	1	426	使用中
1	副		3	3783	使用中
231	232	233	234	235	236

230

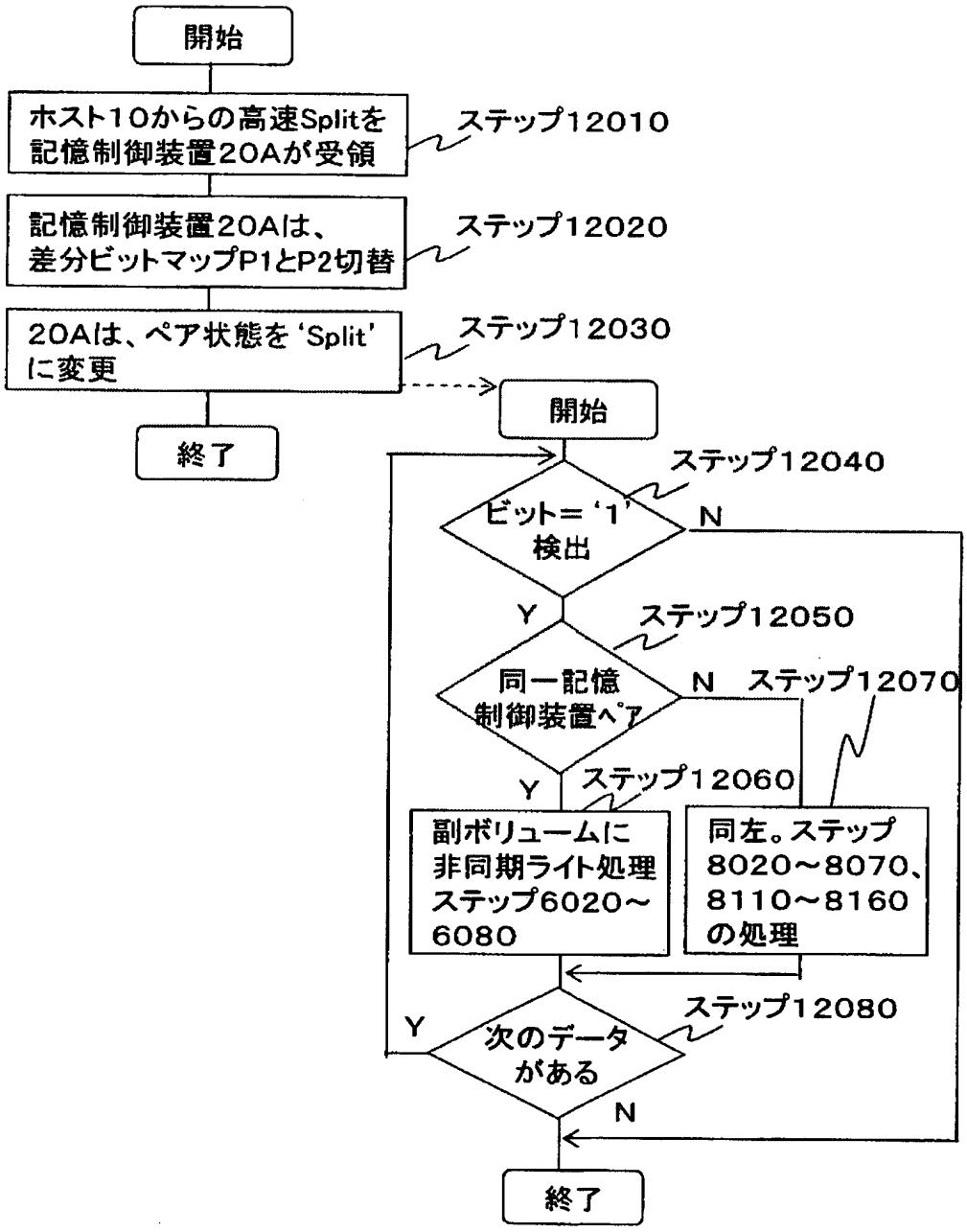
【図１１】

図11



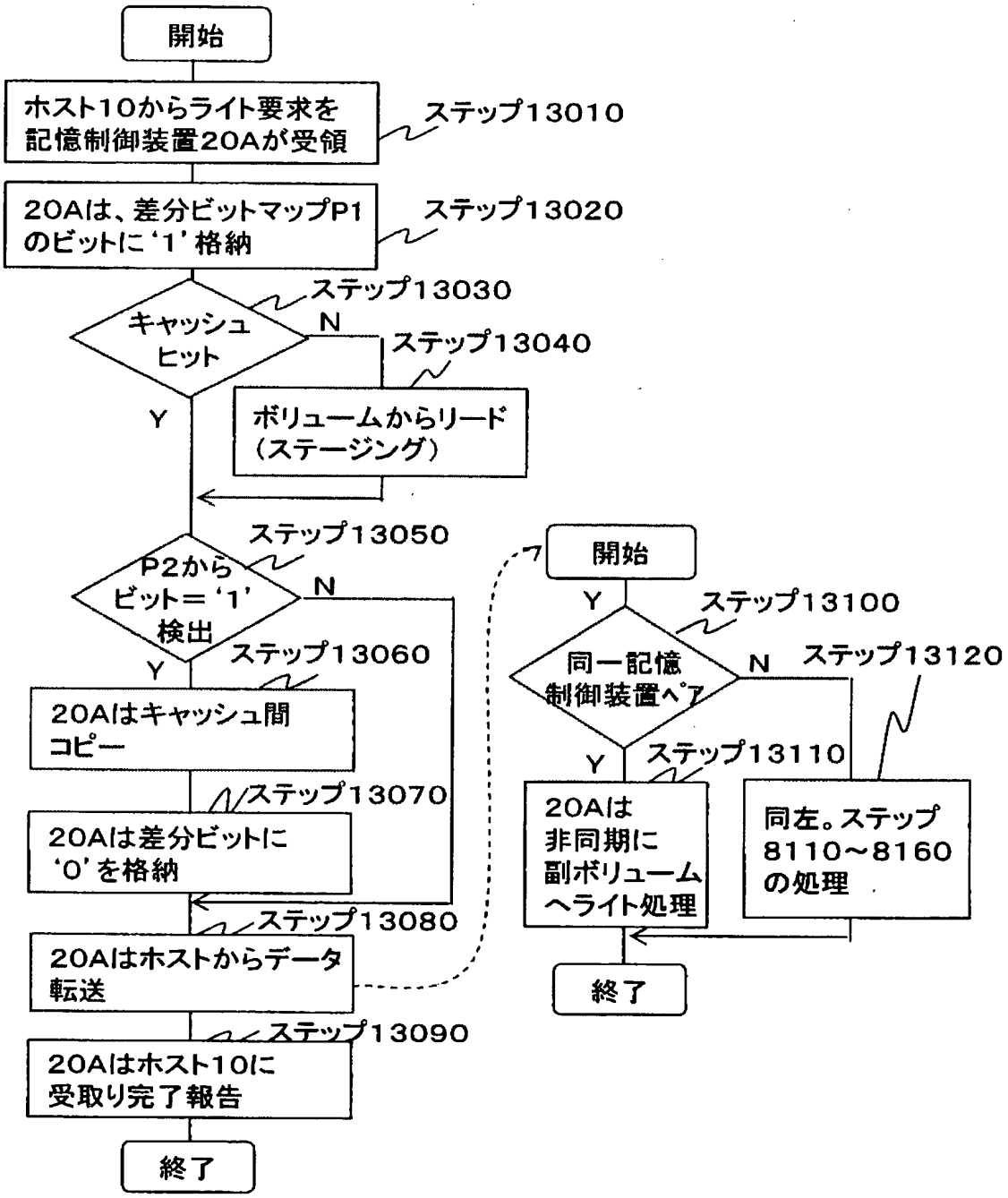
【図１２】

図12


開始
ホスト10からの高速Splitを記憶制御装置20Aが受領
ステップ12010
記憶制御装置20Aは、差分ビットマップP1とP2切替
ステップ12020
20Aは、ペア状態を'Split'に変更
ステップ12030
終了
開始
ステップ12040
ビット＝'1'検出
N
Y
ステップ12050
同一記憶制御装置ペア
N
ステップ12070
Y
ステップ12060
副ボリュームに非同期ライト処理ステップ6020～6080
同左。ステップ8020～8070、8110～8160の処理
ステップ12080
Y
次のデータがある
N
終了

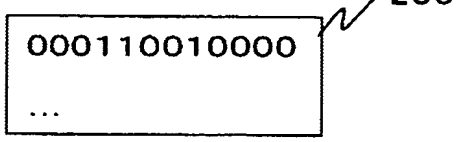
【図１３】

図13



【図１４】

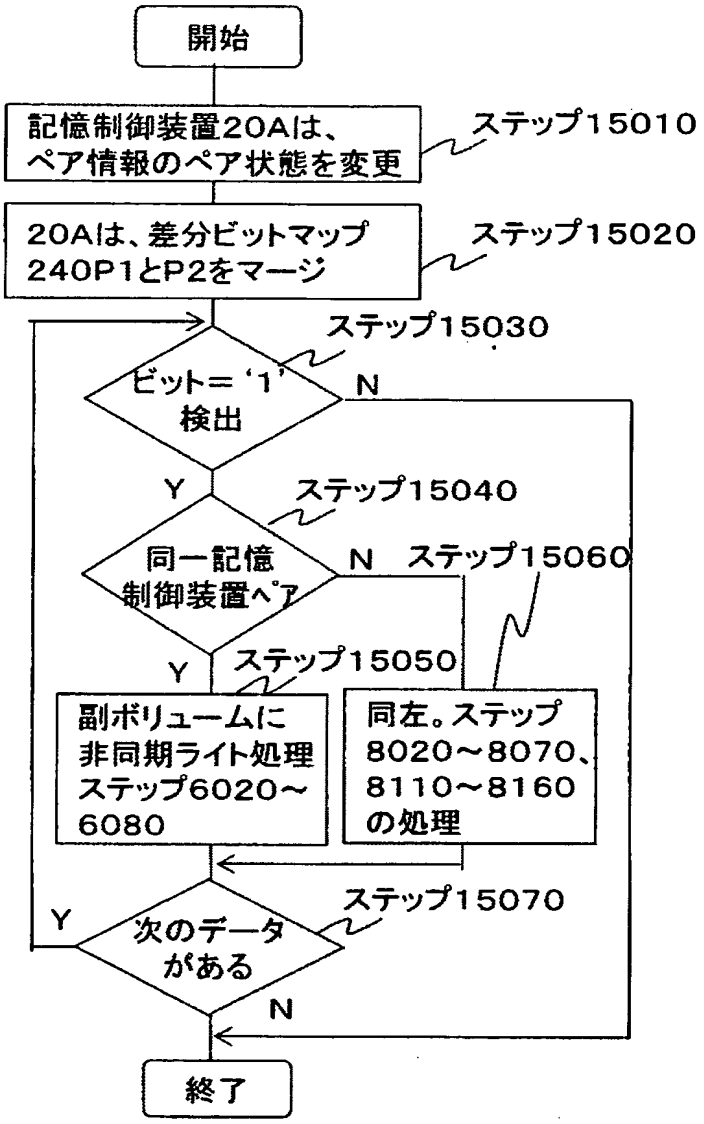
図14



処理優先度情報ビットマップ

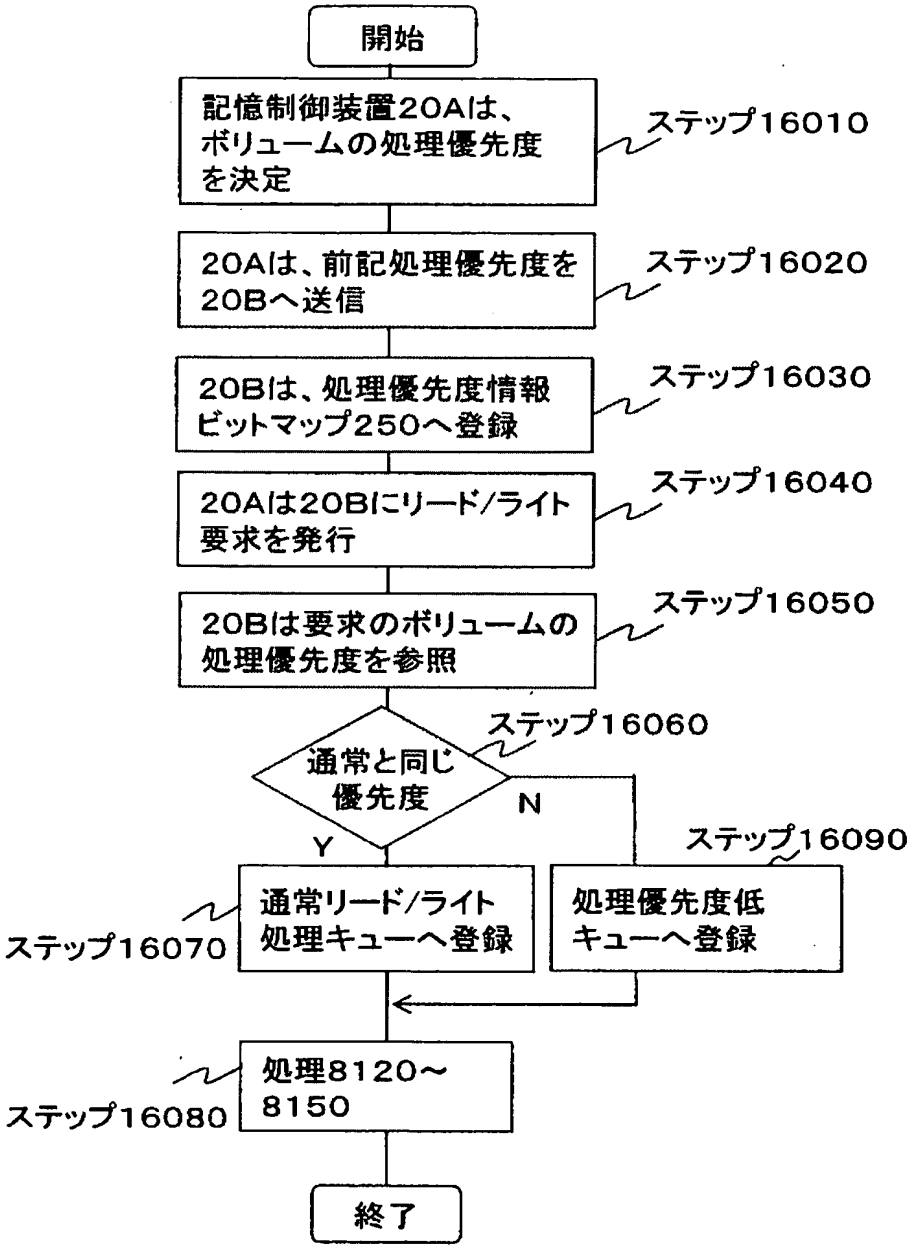
【図１５】

図15


開始
記憶制御装置20Aは、ペア情報のペア状態を変更
ステップ15010
20Aは、差分ビットマップ240P1とP2をマージ
ステップ15020
ステップ15030
ビット＝'1'検出
N
Y
ステップ15040
同一記憶制御装置ペア
N
ステップ15060
Y
ステップ15050
副ボリュームに非同期ライト処理ステップ6020～6080
同左。ステップ8020～8070、8110～8160の処理
ステップ15070
Y
次のデータがある
N
終了

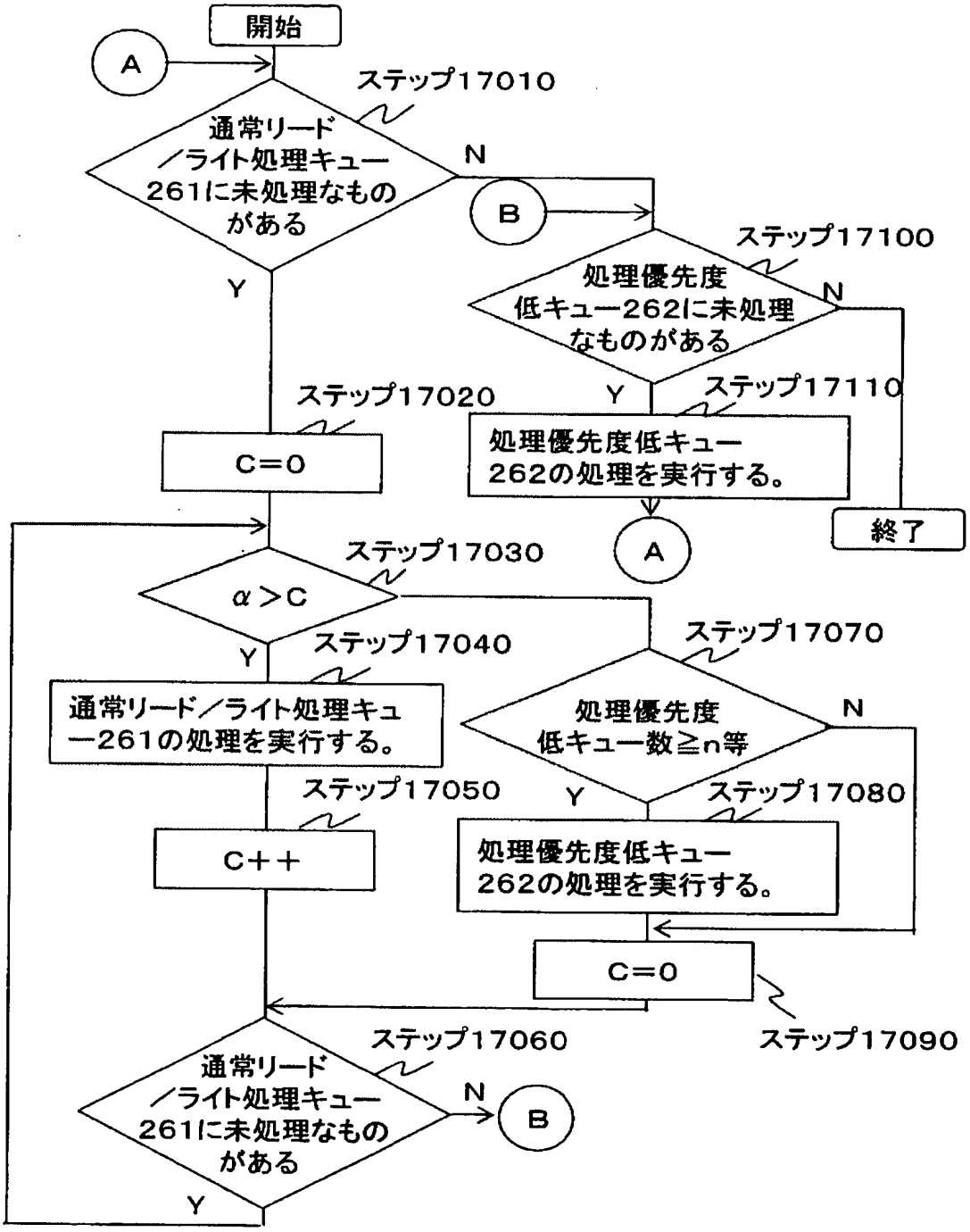
【図１６】

図16


開始
記憶制御装置20Aは、ボリュームの処理優先度を決定
ステップ16010
20Aは、前記処理優先度を20Bへ送信
ステップ16020
20Bは、処理優先度情報ビットマップ250へ登録
ステップ16030
20Aは20Bにリード/ライト要求を発行
ステップ16040
20Bは要求のボリュームの処理優先度を参照
ステップ16050
通常と同じ優先度
ステップ16060
N
Y
ステップ16090
通常リード/ライト処理キューへ登録
ステップ16070
処理優先度低キューへ登録
処理8120～8150
ステップ16080
終了

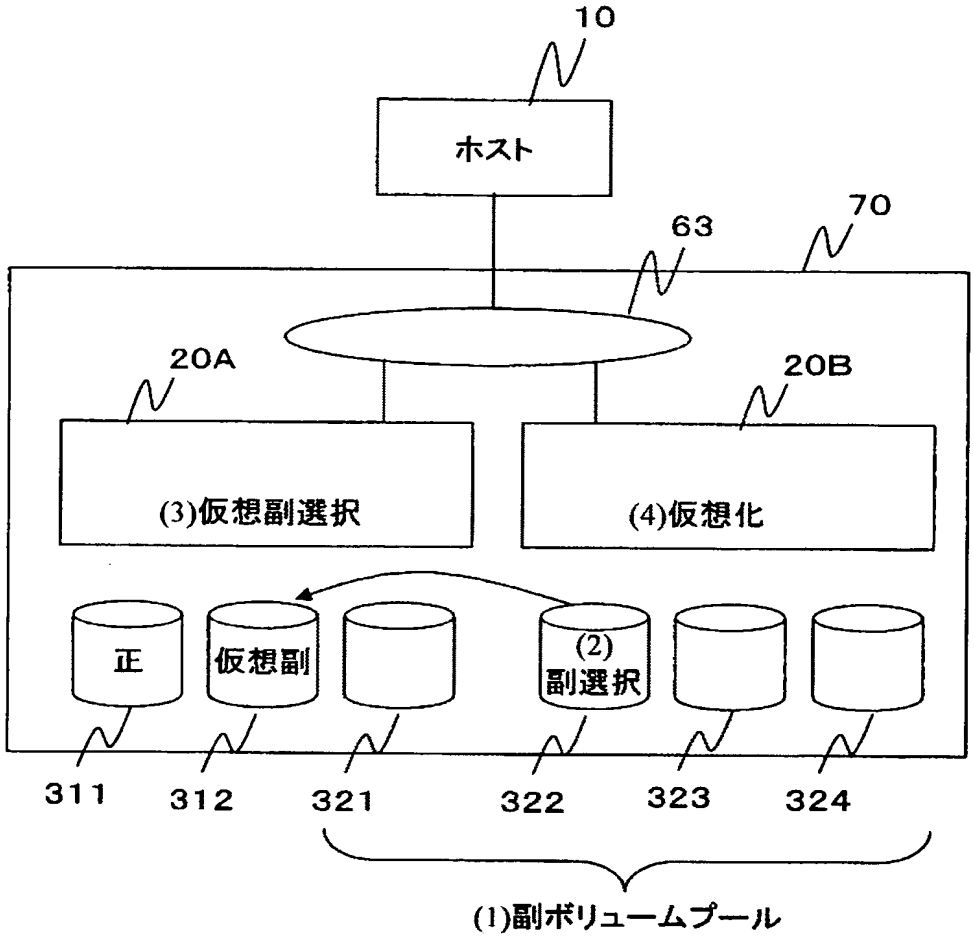
【図１７】

図17


開始
A
ステップ17010
通常リード／ライト処理キュー261に未処理なものがある
N
B
ステップ17100
処理優先度低キュー262に未処理なものがある
N
Y
Y
ステップ17020
C=0
ステップ17110
処理優先度低キュー262の処理を実行する。
A
終了
ステップ17030
α＞C
Y
ステップ17040
通常リード／ライト処理キュー261の処理を実行する。
ステップ17070
処理優先度低キュー数≧n等
N
Y
ステップ17050
C++
ステップ17080
処理優先度低キュー262の処理を実行する。
C=0
ステップ17090
ステップ17060
通常リード／ライト処理キュー261に未処理なものがある
N
B
Y

【図１８】

図18


10
ホスト
63
70
20A
20B
(3)仮想副選択
(4)仮想化
正
仮想副
(2)
副選択
311
312
321
322
323
324
(1)副ボリュームプール

【書類名】要約書
【要約】
【課題】　複数のディスク装置が接続された制御部を複数個有する記憶装置システムにおいて，異なった制御部に接続されたディスク装置内のボリュームにレプリケーションを作成できる技術を提供する。
【解決手段】　一制御部のレプリケーション作成部は，他の制御部に接続されたディスク装置内のボリュームにレプリケーションを作成する場合は，ペア情報に，レプリケーション元のボリューム情報と，前記一制御部におけるレプリケーション先のボリューム情報と，前記他の制御部に関する情報と，を登録し，前記ペア情報に基づいて，レプリケーションを作成するための要求を前記他の制御部へ送信する。
【選択図】　図７

認定・付加情報

特許出願の番号	特願２００３－４０３８６８
受付番号	５０３０１９８９４９０
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１５年１２月　４日

＜認定情報・付加情報＞

【提出日】　平成15年12月 3日

特願２００３－４０３８６８

出　願　人　履　歴　情　報

識別番号　［０００００５１０８］

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所